Київський столичний університет імені Бориса Грінченка
Факультет інформаційних технологій та математики
Кафедра інформаційної та кібернетичної безпеки
імені професора Володимира Бурячка

«Допущено до захисту»
Завідувач кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
кандидат технічних наук, доцент
Складанний П.М.

_____
(підпис)

« ___ » _____ 20__ р.

КВАЛІФІКАЦІЙНА РОБОТА
на здобуття другого (магістерського)
рівня вищої освіти

Спеціальність 125 Кібербезпека та захист інформації

Тема роботи:
ДОСЛІДЖЕННЯ МЕТОДІВ ТА ЗАСОБІВ ПРОТИДІЇ СПАМУ

Виконав

студент групи БІКСм-1-24-1.4.д

Губар Олександр Сергійович

_____
(підпис)

Науковий керівник

к. т. н., доцент

Козачок В.А.

_____
(підпис)

Київ – 2025

Київський столичний університет імені Бориса Грінченка
Факультет інформаційних технологій та математики
Кафедра інформаційної та кібернетичної безпеки
імені професора Володимира Бурячка

Освітньо-кваліфікаційний рівень – магістр
Спеціальність 125 Кібербезпека та захист інформації
Освітня програма 125.00.01 Безпека інформаційних і комунікаційних систем

«Затверджую»
Завідувач кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
кандидат технічних наук, доцент
Складанний П.М.

_____
(підпис)
« ___ » _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

Губару Олександру Сергійовичу

1. Тема роботи: Дослідження методів та засобів протидії СПАМу; керівник Козачок Валерій Анатолійович к.т.н.,доцент, затверджені наказом ректора від «21» серпня 2025 року №482.
2. Термін подання студентом роботи «08» грудня 2025 р.
3. Вихідні дані до роботи:
   - 3.1 науково-технічна та нормативна література з теми дослідження: Наукові публікації з методів виявлення спаму, стандарти кібербезпеки ДСТУ та ISO, RFC для електронної пошти, матеріали з ML та NLP.
   - 3.2 методи: Методи класифікації тексту, NLP-обробка (TF-IDF, токенізація, лематизація), алгоритми ML (Naive Bayes, SVM, Random Forest, нейромережі), порівняльний аналіз точності.
   - 3.3 технології: Python, Scikit-learn, Jupyter Notebook, засоби побудови та оцінювання моделей класифікації.
   - 3.4 алгоритми: TF-IDF для формування ознак, Multinomial Naive Bayes, Support Vector Machine

(linear kernel), Random Forest.

4. Зміст текстової частини роботи (перелік питань, які потрібно розробити):

4.1 Аналіз сучасних методів і засобів виявлення та класифікації спаму.

4.2 Розроблення та побудова моделей машинного навчання для виявлення спаму (TF-IDF, Naive Bayes, SVM, Random Forest)..

4.3 Порівняльна оцінка точності та ефективності моделей спам-класифікації.

5. Перелік графічного матеріалу:

5.1 Презентація доповіді, виконана в Microsoft PowerPoint.

5.2 Схеми обробки текстових даних та структурні діаграми моделей класифікації.

6. Дата видачі завдання «15» лютого 2025р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів підготовки роботи	Термін виконання	Примітка
1.	Уточнення постановки завдання	20.03.2025-28.03.2025	виконано
2.	Аналіз літератури	29.03.2025-04.03.2025	виконано
3.	Обґрунтування вибору рішення	04.03.2025-10.03.2025	виконано
4.	Збір даних	23.09.2025-16.10.2025	виконано
5.	Виконання та оформлення розділу 1.	21.10.2025-01.11.2025	виконано
6.	Виконання та оформлення розділу 2.	12.11.2025-18.11.2025	виконано
7.	Виконання та оформлення розділу 3.	21.11.2025-26.11.2025	виконано
8.	Вступ, висновки, реферат	21.10.2025-26.11.2025	виконано
9.	Апробація роботи на науково-методичному семінарі та/або науково-технічній конференції	26.10.2025	виконано
10.	Оформлення та друк текстової частини роботи	10.12.2025	виконано
11.	Оформлення презентацій	06.12.2025-10.12.2025	виконано
12.	Отримання рецензій	01.12.2025	виконано
13.	Попередній захист роботи	20.11.2025	виконано
14.	Захист в ЕК	16.12.2025-18.12.2025	виконано

Студент _____ Губар Олександр Сергійович
(підпис) (прізвище, ім’я, по батькові)

Науковий керівник _____ Козачок Валерій Анатолійович
(підпис) (прізвище, ім’я, по батькові)

РЕФЕРАТ

Кваліфікаційна робота присвячена технологіям використання методів машинного навчання та обробки природної мови в системах протидії спаму.

Робота складається зі вступу, трьох розділів, що містять 8 рисунків та 3 таблиці, висновків та списку використаних джерел, що містить 31 найменування. Загальний обсяг роботи становить 75 сторінок, з яких 23 сторінки займають ілюстрації і таблиці на окремих аркушах, а також додатки, перелік умовних скорочень та список використаних джерел.

Об'єктом дослідження в роботі є процес виявлення та фільтрації спаму в електронних комунікаційних системах.

Предметом дослідження є методи інтелектуального аналізу даних і NLP-технології для протидії спам-загрозам.

Метою роботи є дослідження, порівняння та систематизація сучасних методів протидії спаму й формування практичних рекомендацій щодо їх застосування.

Для досягнення поставленої мети у роботі:

-проведено аналіз існуючих підходів до виявлення спаму та фішингових повідомлень;

-досліджено особливості застосування методів машинного навчання та обробки природної мови;

-обґрунтовано вибір оптимальних алгоритмів класифікації спаму та сформовано рекомендації щодо їх використання;

Наукова новизна одержаних результатів полягає в систематизації еволюції антиспам-підходів від статистичних моделей до сучасних нейронних архітектур, виявленні їх сильних і слабких сторін та формуванні обґрунтованих рекомендацій щодо вибору оптимальних засобів залежно від типів спам-загроз та вимог організації.

Галузь застосування. Запропоновані підходи можуть бути використані для створення масштабованих систем виявлення спаму з можливістю адаптації під різні типи трафіку та загроз.

Ключові слова: СПАМ, КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, NLP, ФІЛЬТРАЦІЯ ПОВІДОМЛЕНЬ, SVM, RANDOM FOREST.

ЗМІСТ

СПИСОК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

NLP – Natural Language Processing – обробка природної мови

CNN – Convolutional Neural Network – згорткова нейронна мережа

RNN – Recurrent Neural Network – рекурентна нейронна мережа

LSTM – Long Short-Term Memory – тип RNN з механізмом пам’яті

BERT – трансформерна модель

TF-IDF – метод формування ознак тексту

SVM – Support Vector Machine – метод опорних векторів

Random Forest – ансамблевий алгоритм класифікації

k-NN – k-Nearest Neighbors – алгоритм найближчих сусідів

F1-score – Метрика якості класифікації

ВСТУП

Актуальність теми. За останнім десятиліттям масштаби глобальної проблеми спаму значно зросли. Відповідно до свідчень провідних досліджень, щодня у світі надсилається понад триста п'ятдесят мільярдів спам-повідомлень, при цьому спам становить понад вісімдесят п'ять відсотків від усього обсягу електронної пошти. На підприємствах частка небажаних повідомлень у вхідній кореспонденції коливається в межах від 50 до 90 %. Економічні втрати від спаму та пов'язаних з ним фішингових атак для глобальної економіки сягають більш ніж сто сорок мільйонів доларів США щорічно, при цьому витрати компаній на боротьбу зі спамом невпинно зростають.

Суттєвий негативний вплив проявляється не лише у фінансових показниках. Спамерські атаки виступають основним вектором поширення фішингових листів, шкідливого програмного забезпечення та інших кібератак. За даними дослідження, фішингові листи, передані через спам-канали, мають приблизно двадцятивідсотковий рівень успішності; водночас один успішний фішинг нерідко призводить до компрометації критичної інформації та неявних збитків. Малі та середні підприємства особливо вразливі до таких атак, оскільки часто не мають спеціалізованих відділів з кібербезпеки та покладаються на базові засоби захисту. Корпоративна статистика свідчить, що близько вісімдесяти п'яти відсотків компаній розміром до п'ятисот працівників регулярно піддаються спам-атакам та фішинговим спробам.

На сучасному етапі розвитку технологій штучного інтелекту та генеративних мовних моделей проблема спаму набула нової якості. Використання ChatGPT та подібних систем дозволяє зловмисникам генерувати високоякісні, персоналізовані фішингові листи у масштабі, що було неможливо раніше. Дослідження показали, що ChatGPT генеровані фішингові листи демонструють рівень успішності на рівні близько 70 %, порівняно з атаками, написаними людьми, при цьому суттєву перевагу отримують спамери через можливість автоматичного генерування тисяч унікальних варіацій персоналізованого контенту за годину.

На тлі цих викликів пошук новітніх методів та технологій протидії спаму становить невідкладне науково-практичне завдання. Традиційні методи фільтрації, засновані на простому аналізі ключових слів та сигнатурах, давно втратили ефективність проти еволюціонуючих загроз. Кількісна оцінка цієї неспроможності свідчить про те, що системи, які використовують лише текстові фільтри та чорні списки, досягають точності виявлення спаму не більше 90 % з суттєво вищим рівнем помилкових позитивних спрацьовувань. Цей факт вказує на критичну необхідність запровадження більш витончених, адаптивних та інтелектуальних методів аналізу.

Мета роботи. Дослідження та систематизація сучасних методів і засобів протидії спаму на основі аналізу традиційних та інноваційних підходів машинного навчання, глибинного навчання та обробки природної мови, з метою розробки комплексних рекомендацій щодо їх практичного застосування в системах електронної пошти та комунікацій.

Для досягнення поставленої мети в роботі розв'язуються такі завдання:

1. провести комплексний аналіз видів спаму та фішингових атак, визначити їх особливості та механізми поширення.
2. дослідити традиційні методи виявлення спаму, включаючи статистичні підходи та класичні алгоритми машинного навчання.
3. вивчити сучасні підходи, засновані на глибинному навчанні, включаючи згорткові нейронні мережі (CNN), рекурентні нейронні мережі (LSTM) та трансформерні архітектури (BERT).
4. розглянути методи обробки природної мови для вилучення та нормалізації текстових ознак.
5. провести порівняльний аналіз ефективності різних методів та визначити їх оптимальні сфери застосування.
6. окреслити перспективні напрями розвитку антиспам-технологій у контексті розвитку штучного інтелекту та нових загроз.

Об'єктом дослідження. Методи, алгоритми та технологічні рішення, спрямовані на виявлення та фільтрацію спаму в системах електронної комунікації.

Предмет дослідження. Методи інтелектуального аналізу даних та обробки природної мови як засоби протидії спам-загрозам у системах електронної пошти та месенджерів.

Методи дослідження. У роботі використано теоретичні та емпіричні методи: системний аналіз літератури, порівняльний аналіз методів, аналіз статистичних даних щодо ефективності різних підходів на відкритих датасетах (Enron, SpamAssassin, Lingspam, SPMDC), метод синтезу результатів дослідження.

Наукова новизна одержаних результатів полягає у систематизації та комплексному аналізі еволюції методів протидії спаму від традиційних статистичних підходів до сучасних архітектур глибинного навчання, виявленні їх сильних та слабких сторін, а також у формуванні науково обґрунтованих рекомендацій щодо вибору оптимальних методів залежно від специфіки організації та типів загроз, яким вона піддається.

Теоретичне та практичне значення отриманих результатів полягає в тому, що матеріали роботи можуть слугувати основою для розробки та впровадження ефективних систем протидії спаму в організаціях різних розмірів, від малих підприємств до великих корпорацій, а також використовуватися як навчальний матеріал у підготовці спеціалістів у сфері кібербезпеки та інформаційних технологій.

Галузь застосування. Результати роботи можуть бути використані для розробки та впровадження ефективних систем протидії спаму на підприємствах малого та середнього бізнесу, в установах державного та приватного сектору, а також в інтегрованих рішеннях для великих корпоративних мереж. Матеріали дослідження можуть служити основою для удосконалення існуючих антиспам-фільтрів та розробки нових методологій захисту електронної комунікації. Крім того, результати можуть бути використані як навчальний матеріал у навчальному процесі при підготовці фахівців у сфері інформаційної безпеки, кібербезпеки та комп'ютерни

х наук.

Апробація результатів дипломної роботи. Основні положення роботи представлені та апробовані в таких виданнях:

у збірнику тез Студентської наукової конференції «Безпека інформаційно-комунікаційних систем» (БІКС’2025), яка пройшла 26 жовтня 2025 року на базі Київського столичного університету імені Бориса Грінченка.

Розділ 1. ТЕОРЕТИЧНІ ОСНОВИ ТА АНАЛІЗ СУЧАСНИХ МЕТОДІВ ПРОТИДІЇ СПАМУ

1.1 Проблематика спаму та класифікація сучасних загроз

Спам є однією з ключових проблем сучасного інформаційного суспільства, оскільки суттєво впливає як на якість електронних комунікацій, так і на стійкість інформаційної інфраструктури в цілому. Зростання обсягів небажаної кореспонденції призводить до значних втрат часу користувачів, нераціонального використання обчислювальних ресурсів і пропускної здатності мереж, а також формує сприятливе середовище для реалізації різноманітних кіберзагроз [1]. За оцінками дослідників, близько сорока відсотків електронного листування становить спам, що відповідає приблизно 15,4 мільярда повідомлень на добу, тоді як сумарні економічні втрати користувачів інтернету оцінюються на рівні сотень мільйонів доларів щороку. Такі показники свідчать про перетворення спаму з локальної незручності на глобальний фактор ризику, який вимагає системного вивчення та розробки ефективних механізмів протидії [2].

Окрім безпосередніх незручностей для кінцевих користувачів, спам зумовлює низку технічних та організаційних проблем для провайдерів послуг і корпоративних мереж. Масові розсилки призводять до перевантаження поштових серверів, неефективного використання дискового простору, зниження продуктивності мережевої інфраструктури та зростання витрат на обслуговування та модернізацію систем фільтрації [3]. У сукупності це ускладнює підтримання належного рівня якості сервісів та змушує організації інвестувати додаткові ресурси у підтримку працездатності поштових систем.

Разом з тим спам виступає не лише джерелом перевантаження інфраструктури, а й одним з основних каналів доставки шкідливого контенту. Через спам-повідомлення розповсюджуються фішингові листи, шкідливе програмне забезпечення, а також реалізуються сценарії соціальної інженерії, спрямовані на

обман користувачів та отримання несанкціонованого доступу до конфіденційних даних. У таких умовах спам слід розглядати як базовий інструмент для побудови більш складних атак, які можуть призводити до компрометації облікових записів, порушення цілісності інформації та дестабілізації роботи інформаційних систем [4, 5].

Важливою особливістю спаму є постійна еволюція способів його створення та розповсюдження. Якщо на початкових етапах протидії було достатньо блокувати повідомлення з певних адрес або фільтрувати їх за фіксованими ключовими словами, то згодом спамери адаптувалися до таких підходів. Нині широко використовуються випадкові або динамічно згенеровані адреси відправників, варіативні теми листів, додавання випадкових символів чи фрагментів тексту, а також маскування змісту з метою ускладнення аналізу [6, 38, 39]. Такі прийоми обфускації суттєво знижують результативність традиційних фільтрів, побудованих на статичних експертних правилах, і вимагають їхнього постійного доопрацювання.

Для глибшого розуміння природи явища спаму доцільно розглядати його у розрізі кількох класифікаційних ознак. За каналами розповсюдження спам може надходити через електронну пошту, SMS-повідомлення, системи миттєвого обміну повідомленнями, соціальні мережі, коментарі на веб-ресурсах та інші платформи електронних комунікацій. Змістовно виділяють, зокрема, комерційний спам, орієнтований на рекламу товарів і послуг; фішингові повідомлення, спрямовані на викрадення конфіденційної інформації; шкідливий спам із посиланнями або вкладеннями, що містять шкідливе програмне забезпечення; соціально-інженерний спам, який експлуатує психологічні вразливості користувачів; а також політичний спам, пов'язаний із поширенням пропаганди та дезінформації [1]. Така класифікація дозволяє чіткіше визначати цілі зловмисників та добирати адекватні методи протидії.

З позицій впливу на безпеку спам доцільно розглядати як багатовекторну загрозу. У контексті інформаційної безпеки він виступає транспортним середовищем для атак, спрямованих на отримання облікових даних, порушення конфіденційності чи цілісності інформації та встановлення шкідливого програмного забезпечення. З

технічної точки зору масові розсилки можуть провокувати перевантаження поштових серверів та окремих сегментів мережі, що у крайніх випадках призводить до відмови в обслуговуванні та недоступності критичних сервісів [3, 7]. Для організацій це означає одночасне зростання оперативних ризиків та потреби у додаткових механізмах контролю і моніторингу трафіку.

Серйозним викликом є й завдання виявлення та класифікації спаму в умовах постійної адаптації зловмисників. Жорстко задані правила та фільтри на основі фіксованих сигнатур виявилися недостатньо гнучкими, оскільки вимагають безперервного ручного оновлення та не встигають за темпами змін у тактиках спамерів [8]. Це зумовило перехід до використання методів машинного навчання, які навчаються на історичних даних і здатні автоматично підлаштовуватися під нові варіанти спам-повідомлень. У результаті моделі класифікації почали враховувати не лише окремі ключові слова, а й контекст, структуру повідомлення, поведінкові характеристики відправників та інші ознаки.

Додаткові труднощі породжує багатомовний характер спаму та потреба в обробці великих обсягів даних у режимі, наближеному до реального часу. Системи фільтрації мають коректно працювати з повідомленнями різними мовами, враховуючи їхні лінгвістичні особливості, а також забезпечувати високу пропускну здатність без помітного впливу на продуктивність поштових сервісів. Паралельно відбувається інтеграція спаму з іншими формами кіберзагроз, зокрема цільовими фішинговими атаками (spear phishing), орієнтованими на конкретних користувачів чи організації. Це підвищує вимоги до точності класифікації та зменшення кількості помилкових спрацьовувань [9].

Суттєву роль у сучасних спам-кампаніях відіграють автоматизовані інструменти, передусім ботнети, що здатні здійснювати масові розсилки за короткі часові інтервали. Зловмисники активно використовують і легітимні сервіси — зокрема хмарні сховища та авторитетні домени — для розміщення спам-контенту, надсилаючи користувачам лише гіперпосилання з мінімальним текстовим

супроводом. Такий підхід ускладнює виявлення загроз на основі аналізу лише текстової частини повідомлення та вимагає залучення додаткових механізмів перевірки репутації доменів та аналізу посилань [3, 10].

У сукупності ці фактори формують складний, динамічний та такий, що постійно еволюціонує ландшафт спам-загроз. Це обумовлює необхідність безперервного вдосконалення методів протидії, впровадження інтелектуальних систем аналізу даних та розробки науково обґрунтованих підходів до захисту інформаційних систем від небажаного трафіку [11, 33, 34].

1.2 Огляд існуючих методів і алгоритмів виявлення та фільтрації спаму

Сигнатурні методи посідають одну з ключових позицій серед класичних підходів до виявлення спаму, оскільки ґрунтуються на пошуку наперед відомих шаблонів, характерних послідовностей символів та специфічних ключових слів, типових для спам-повідомлень. У таких системах формується база сигнатур, яка постійно поповнюється на основі вже ідентифікованих зразків спаму, після чого кожне нове повідомлення порівнюється з цими еталонами. Це забезпечує високу швидкість обробки та точність виявлення раніше відомих шаблонів без потреби у навчанні складних моделей. На практиці системи на кшталт SpamAssassin застосовують сотні окремих тестів для комплексної оцінки вхідної кореспонденції; пропускна здатність таких рішень може сягати близько 10 000 повідомлень за секунду при низьких обчислювальних витратах [1, 8]. Водночас принципова вада сигнатурних методів полягає в тому, що вони майже не реагують на нові або модифіковані форми спаму, особливо за активного використання технік обфускації, зміни структури листів та поліморфних варіацій контенту, які не збігаються з уже відомими сигнатурами.

Фільтри на основі чорних списків (blacklist-based filtering) реалізують інший фундаментальний підхід, який зазвичай використовується у поєднанні з іншими методами. Його сутність полягає у веденні та регулярному оновленні великих баз

даних підозрілих електронних адрес, доменів та IP-адрес, з яких систематично надходить спам-трафік [3, 6]. Перевірка відправника за таким списком є обчислювально простою операцією, що забезпечує дуже високу пропускну здатність і дозволяє ефективно відсіювати значну частину небажаних повідомлень ще на початкових етапах обробки. Саме простота реалізації та низька вартість обслуговування зробили цей підхід популярним як засіб попередньої фільтрації великих обсягів пошти. Однак ефективність цього методу безпосередньо залежить від актуальності бази: при появі спаму з нових або змінених джерел результативність різко знижується. Спамери активно змінюють IP-адреси, задіюють тимчасові домени та проксі-сервери, унаслідок чого ефективність класичних чорних списків за умов появи нових джерел може падати до 70–80%, що робить цей підхід недостатнім як єдине рішення.

Статистичні методи фільтрації, зокрема наївний байєсівський класифікатор та підходи на основі TF-IDF (Term Frequency–Inverse Document Frequency), використовують імовірнісні моделі для кількісної оцінки ймовірності належності повідомлення до класу спаму на основі статистичних характеристик його текстового вмісту. Наївний Байєс завдяки простоті, невисоким вимогам до ресурсів та добрій узагальнювальній здатності стабільно демонструє точність виявлення спаму на рівні 95–97% на різних наборах даних [12]. TF-IDF дозволяє формувати векторні подання документів, виділяючи терміни, що є статистично значущими саме для спам-повідомлень; за даними низки досліджень, точність класифікації за таким підходом може досягати 97–98% [8]. Водночас імовірнісні методи виявляють підвищену чутливість до штучної зміни розподілів слів, маніпуляцій з лексикою та мовних патернів, які спамери цілеспрямовано змінюють для обходу фільтрів. Знижується ефективність таких алгоритмів і в ситуаціях, коли потрібно аналізувати зображення з вбудованим текстом, мультимедійний контент чи тексти кількома мовами одночасно.

Для узагальнення ключових характеристик різних підходів до фільтрації спаму доцільно розглянути їх у вигляді порівняльної таблиці, див. табл. 1.1.

Таблиця 1.1

Порівняльна характеристика методів виявлення спаму.

Метод	**Точність (%)**	**Пропускна здатність**	**Обчислювальні витрати**	**Адаптивність**	Основні обмеження
Сигнатурні методи	90–94	≈ 10 000 повідомлень/с	Мінімальні	Низька	Неефективні проти нових та модифікованих атак
Чорні списки	70–80	10 000+ повідомлень/с	Мінімальні	Дуже низька	Критична залежність від актуальності баз
Наївний Баєс	95–97	Висока	Низькі	Середня	Чутливість до змін мовних патернів
TF-IDF	97–98	Висока	Низькі	Середня	Вразливість до обфускації та штучних змін лексики
SVM	98–98,5	Середня	Середні	Висока	Потреба у регулярному перенавчанні
Random Forest	99–99,91	Середня	Середні	Висока	Ресурсоємність на дуже великих наборах даних
CNN	97–97,5	Середня	Високі	Дуже висока	Значні обчислювальні витрати

RNN / LSTM	96–97	Низька	Дуже високі	Дуже висока	Повільне навчання, складність реалізації
BERT	98–98,67	Низька	Дуже високі	Дуже висока	Вкрай високі вимоги до апаратних ресурсів
Гібридні системи	98–99+	Середня	Середні–високі	Дуже висока	Складність інтеграції та супроводу

Методи машинного навчання, зокрема машини опорних векторів (SVM), алгоритми випадкового лісу (Random Forest), k-найближчих сусідів (k-NN) та інші класичні моделі, демонструють відчутне зростання точності порівняно з базовими статистичними підходами. Їхньою перевагою є здатність виявляти нелінійні залежності та приховані закономірності в наборі ознак, що описують повідомлення. Random Forest послідовно показує одні з найкращих результатів серед традиційних методів машинного навчання, досягаючи точності до 99,91% при низькому рівні хибнопозитивних спрацьовувань (близько 4,13%), що є критично важливим для систем, де помилка класифікації має значні наслідки [13]. SVM забезпечує точність на рівні близько 98,5% та добре масштабується на великі вибірки, зберігаючи стабільність результатів за умов зростання обсягів спам-трафіку [14]. Разом з тим, усі ці моделі потребують регулярного оновлення та перенавчання на нових даних, щоб залишатися релевантними в умовах постійної змінюваності тактик спамерів.

Глибинне навчання що базується на багатошарових штучних нейронних мережах, являє собою наступний етап розвитку антиспам-технологій. Такі моделі не покладаються виключно на явні, вручну сконструйовані ознаки, а самостійно формують багаторівневі представлення тексту. CNN спочатку розроблені для аналізу зображень, виявили високу ефективність і в задачах обробки текстових

послідовностей завдяки здатності виявляти локальні патерни та стійкі фразові конструкції; у більшості досліджень вони демонструють середню точність близько 97,5% на стандартних наборах даних. Рекурентні архітектури, зокрема мережі LSTM, краще відображають послідовний характер тексту та контекстні залежності між словами, що дозволяє досягати точності понад 97%, але за рахунок значно більшого часу навчання та підвищених обчислювальних витрат порівняно з CNN [15].

BERT використовує механізм самоуваги та двобічний контекст, досягаючи точності до 98,67% та демонструючи одну з найкращих здатностей до узагальнення та роботи з різноманітними типами текстів [9]. Головними обмеженнями глибинних моделей є потреба у потужному апаратному забезпеченні, значні часові витрати на навчання та складність інтерпретації результатів.

Таблиця 1.2

Еволюція методів фільтрації спаму (1995–2025 рр.)

Період	**Метод**	**Принцип роботи**	**Переваги**	**Недоліки**	**Приклади інструментів**
1995–2002	Чорні та білі списки (RBL)	Блокування електронних листів за IP-адресами або доменами (DNSBL)	Простота та висока швидкість	Легко обходяться ботнетами й динамічними IP	MAPS RBL, Spamhaus, SORBS

1998–2005	Сигнатурні фільтри	Пошук шаблонів і ключових фраз	Висока точність для відомих типів спаму	Потребує оновлення правил	SpamAssassin (до 2004), ранній Outlook
2002–2010	Байєсівські фільтри	Ймовірність спаму за словами	Самонавчання, адаптивність	Потребує великого корпусу даних	SpamAssassin, Thunderbird
2004–2015	Класичне машинне навчання	SVM, Random Forest, логістична регресія	Висока точність	Складність інженерії ознак	Gmail (після 2010), Yahoo Mail, Barracuda
2015–2020	Глибокі нейромережі (RNN, LSTM)	Аналіз послідовностей слів та семантики	Розуміння сенсу, краще обходить спам	Потребує великих ресурсів	Gmail, TensorFlow-моделі
2018–2023	Трансформери (BERT)	Контекстне моделювання листа	Дуже висока точність (~99,9%)	Великі моделі, дорогий інференс	Gmail BERT, Microsoft Defender

2023–2025+	LLM-моделі	Аналіз тексту, вкладень, поведінки	Майже людський рівень розуміння	Висока вартість та питання приватності	Google Gemini, OpenAI Outlook, Cloudflare 2025

Гібридні підходи та ансамблеві системи, що поєднують кілька методів у межах єдиного рішення, розглядаються як один із найперспективніших напрямів розвитку антиспам-технологій. Поєднання CNN з RNN або LSTM дозволяє одночасно враховувати локальні патерни тексту та довгострокові контекстні залежності, тоді як додавання класичних моделей машинного навчання на верхньому рівні може покращувати підсумкову класифікацію [16]. Ансамблеві стратегії, що агрегують прогнози кількох незалежних моделей (наприклад, ансамбль із Random Forest, SVM та нейронних мереж), як правило, знижують кількість як хибнопозитивних, так і хибнонегативних рішень за рахунок усереднення або зваженого голосування [17]. Інтеграція сигнатурних, статистичних, класичних машинних та глибинних моделей у рамках єдиної багаторівневої системи дає змогу досягти балансу між точністю, швидкодією та масштабованістю. Водночас такі системи є більш складними з точки зору проєктування, інтеграції та подальшого супроводу, що висуває підвищені вимоги до організацій, які планують їх впроваджувати.

1.3. Сучасні виклики у боротьбі зі спамом

Сучасні виклики у боротьбі зі спамом формують складний, багатоаспектний комплекс проблем, що є наслідком постійної еволюції тактик спамерів та їхньої швидкої адаптації до нових технологій захисту. На концептуальному рівні йдеться про перехід від примітивних масових розсилок до цілеспрямованих, персоналізованих атак, підсилених можливостями штучного інтелекту та машинного навчання [1].

Протягом останніх років відбулася якісна трансформація природи спам-загроз: методи фільтрації, які розроблялися та вдосконалювалися протягом двох десятиліть, виявилися недостатніми для протидії новим формам атак. Ситуацію додатково ускладнює те, що спамери більше не обмежуються простими текстовими повідомленнями чи очевидними маркерами компрометації, а будують багатовекторні кампанії, спрямовані одночасно на технічні системи та людський фактор із використанням складних сценаріїв соціальної інженерії.

Одним із ключових викликів сучасного етапу є використання зловмисниками генеративних моделей штучного інтелекту для автоматизованого створення переконливих та персоналізованих спам- і фішингових повідомлень. Моделі на кшталт GPT-4 та його похідних, а також нелегальні сервіси типу WormGPT та FraudGPT дають змогу автоматизувати збір відкритих персональних даних про жертву, формувати гіперперсоналізовані тексти, генерувати граматично коректний та стилістично природний контент, а також створювати тисячі унікальних варіантів атак за лічені хвилини. У дослідженні IBM продемонстровано, що штучний інтелект зміг підготувати фішингову кампанію зі співставною ефективністю всього за кілька запитів та приблизно п’ять хвилин, тоді як команді експертів-людей для створення аналогічної кампанії знадобилося близько шістнадцяти годин роботи. Аналітики SentinelOne зафіксували зростання кількості фішингових атак, пов’язаних із використанням генеративного ШІ, на 1265% протягом одного року, що свідчить про різкий зсув у бік автоматизованої злочинної діяльності. За даними досліджень Гарвардського університету, близько 60% отримувачів переходять за посиланнями у фішингових листах, створених ШІ, при тому що витрати спамерів на підготовку таких кампаній скорочуються до 5% від традиційних. ФБР США офіційно попереджає, що кримінальні угруповання активно використовують штучний інтелект для організації високоцільових фішингових кампаній, які призводять до значних фінансових втрат, репутаційних збитків та витоків конфіденційної інформації [5, 19].

Проблематика загострюється через поєднання спаму з мультимедійними дипфейками та синтезованим аудіо- й відеоконтентом. Сучасні технології дають змогу створювати надзвичайно переконливі відеозвернення й голосові повідомлення, які імітують реальних керівників або довірених осіб [4, 19]. Резонансний випадок на початку 2024 року, коли підроблене за допомогою ШІ відео фінансового директора спонукало співробітника авторизувати переказ 25 млн доларів США на рахунки зловмисників, показав практичну критичність таких загроз. Згідно зі звітом IBM про вартість порушень безпеки даних за 2024 рік, середня вартість інциденту, пов’язаного з фішингом, становить 4,88 млн доларів. Близько 64% американських компаній у 2024 році зіткнулися з атаками типу Business Email Compromise (BEC), при цьому середні прямі втрати на одну атаку сягають 150 тис. доларів.

Окрему групу сучасних викликів становлять evasion-атаки, спрямовані на цілеспрямований обхід уже натренованих моделей машинного та глибинного навчання. На відміну від атак отруєння даних, які впливають на процес навчання, evasion-атаки реалізуються на етапі експлуатації моделі: зловмисник модифікує вхідні дані так, щоб змусити класифікатор видати хибне рішення. У сфері спаму це досягається шляхом внесення мінімальних, майже непомітних змін у текст: заміни літер на схожі символи, застосування альтернативних орфографічних варіантів, перебудови фраз, використання синонімів та зміни розмітки [20, 21]. Такі адверсаріальні модифікації часто залишаються невидимими для користувачів, але виявляються достатніми для того, щоб обійти фільтри, орієнтовані на статистичні чи лексичні особливості тексту. Методи адверсаріального тренування, коли модель додатково навчається на штучно спотворених прикладах, підвищують стійкість до таких атак, однак потребують постійного оновлення та значних обчислювальних ресурсів.

Суттєвою перешкодою для побудови ефективних антиспам-систем залишається дефіцит якісних тренувальних даних, особливо в умовах жорстких вимог до приватності та конфіденційності. Дослідження, представлені на конференції ICEB

2022, показали, що значна частина організацій та користувачів не готові передавати приклади реальних спам-повідомлень у хмарні сервіси через побоювання витоку чутливої інформації. Це обмежує можливості для створення репрезентативних, актуальних та доменно-специфічних наборів даних, необхідних для тренування персоналізованих моделей фільтрації. Окремі роботи пропонують використання блокчейн-технологій та токен-орієнтованих механізмів стимулювання як інструментів анонімізованого збору даних про спам, однак впровадження таких рішень на практиці потребує суттєвих інвестицій та організаційних змін [22, 23].

Ще одним принциповим викликом є необхідність одночасно забезпечити високу чутливість систем до спаму та наднизький рівень хибнопозитивних спрацьовувань. Підвищення чутливості моделей природно збільшує ризик помилкового блокування легітимних повідомлень, що для бізнесу може означати втрату важливих контрактів, збої в логістичних ланцюгах чи порушення регуляторних вимог. Дослідження у галузі електроніки та прикладних наук показують, що ансамблеві підходи, які поєднують кілька різних моделей, дозволяють досягати кращого компромісу між повнотою виявлення та рівнем хибнопозитивів. Водночас такі рішення ускладнюють архітектуру систем, потребують додаткових обчислювальних ресурсів та ретельного моніторингу продуктивності. Питання масштабованості також є критичним: сучасні поштові платформи змушені обробляти мільйони повідомлень за секунду, інтегруючи при цьому складні алгоритми аналізу в реальному часі.

Багатомовність та культурна різноманітність спам-контенту додають ще один рівень складності. Більшість промислових систем фільтрації історично орієнтовані на англомовний контент та західні культурні контексти, тоді як сучасні спам-кампанії активно використовують локальні мови, діалекти, архаїзми та культурно специфічні вирази. Широко використовується змішання мов у межах одного повідомлення, що ускладнює роботу моделей, навчених на одномовних корпусах. Гібридні системи, які поєднують сигнатурні, статистичні та нейронні підходи, демонструють кращі

результати в таких умовах, однак потребують постійної адаптації до нових мовних та культурних патернів [9, 24].

Висока адаптивність спамерів та розробників шкідливого ПЗ виявляється не лише в зміні змісту повідомлень, а й у систематичному вивченні наукових публікацій та промислових звітів з кібербезпеки. Зловмисники аналізують оприлюднені методики виявлення спаму для ідентифікації потенційних слабких місць і починають експлуатувати їх ще до того, як захисні рішення отримують масове розповсюдження. Однією з найбільш витончених тактик є використання поліморфізму, коли генеруються тисячі варіантів одного й того самого повідомлення з незначними відмінностями в тексті, структурі HTML, вкладеннях чи метаданих. У поєднанні з генеративними моделями це дає змогу створювати динамічні спам-кампанії, які змінюють свою поведінку залежно від характеристик цільової системи фільтрації [20, 21].

Додатковий вимір складності пов’язаний із проблемою інтерпретованості рішень, що приймаються системами глибинного навчання. Якщо у випадку класичних алгоритмів можна відносно просто з’ясувати, які ознаки призвели до класифікації листа як спаму, то нейронні мережі діють як «чорні скриньки». Це ускладнює пояснення користувачам причин блокування їхніх легітимних повідомлень, створює труднощі для внутрішнього аудиту та перевірки систем на предмет прихованих упереджень, а також викликає додаткові запитання з боку регуляторних органів. Проблема прозорості рішень стає особливо гострою в контексті вимог до відповідального використання ШІ та довіри до автоматизованих систем [1, 17] .

Суттєвої ваги набувають етичні та правові аспекти використання персональних даних у процесі побудови антиспам-рішень. Найефективніші моделі нерідко ґрунтуються на аналізі великих масивів реальної кореспонденції, що містить чутливу інформацію — фінансові, медичні, юридичні чи приватні відомості. Такі дані охоплюються вимогами міжнародних та національних регуляторних актів, зокрема GDPR, CCPA та відповідного національного законодавства [3, 25]. Порушення

принципів мінімізації даних, прозорості обробки та інформованої згоди створює як юридичні ризики для організацій, так і додаткову поверхню атаки: у разі компрометації тренувальних наборів зловмисники можуть отримати доступ до масивів конфіденційної інформації.

Окремий структурний виклик стосується дефіциту кваліфікованих фахівців, здатних проєктувати, розгортати та підтримувати сучасні антиспам-системи. Для цього потрібна комбінація компетентностей у галузі машинного навчання, кібербезпеки, архітектури високонавантажених систем та управління інформаційною безпекою. Багато малих і середніх організацій не мають ресурсу для формування власних команд такого рівня й змушені покладатися на типові хмарні сервіси, які не завжди враховують їхню специфіку. Динамічний розвиток технологій призводить до швидкого старіння знань, що обумовлює потребу в постійному професійному навчанні та підвищенні кваліфікації, яке є дорогим та ресурсомістким [11, 26].

Новий клас загроз формується навколо атак «data poisoning», спрямованих на отруєння тренувальних наборів даних. Зловмисник цілеспрямовано додає до навчальних вибірок спотворені або підроблені приклади, наприклад, маркуючи спам як легітимну кореспонденцію чи навпаки. Навіть відносно невелика частка таких даних у загальному обсязі може призвести до помітного погіршення якості моделі та збільшення кількості помилкових рішень. Захист від атак отруєння потребує складних процедур валідації даних, застосування робастних методів навчання та постійного моніторингу поведінки моделей у продуктивному середовищі [20, 21]. Це збільшує вартість розробки та експлуатації систем фільтрації спаму, але стає необхідною умовою їхньої надійності.

Узагальнюючи, сучасні виклики у боротьбі зі спамом не можуть бути розв’язані за допомогою одного універсального технічного засобу. Потрібен багаторівневий підхід, що поєднує технологічні інновації, ефективне управління даними, нормативно-правові механізми, освітні ініціативи та міжнародну координацію. Швидкість еволюції спам-загроз часто перевищує темпи оновлення нормативної бази,

що фактично підтримує «гонку озброєнь» між захисними технологіями та зловмисниками. Тому сталий успіх у протидії спаму залежатиме не лише від подальшого розвитку методів машинного та глибинного навчання, а й від здатності суспільства вибудувати збалансовані підходи до регулювання, які одночасно враховують вимоги безпеки, права людини та принципи прозорості й підзвітності.

Висновки до першого розділу

У межах першого розділу було показано, що спам перетворився з другорядної незручності для користувачів на стійкий фактор ризику для сучасних інформаційно-комунікаційних систем. Він одночасно створює надмірне навантаження на поштову інфраструктуру, погіршує якість сервісів та слугує основним каналом доставки фішингових повідомлень, шкідливого програмного забезпечення та елементів соціальної інженерії. Детальна класифікація спаму за каналами розповсюдження та за змістом дала змогу чітко відокремити комерційний, шкідливий, фішинговий, соціально-інженерний і політичний спам та показала, що кожен із цих різновидів має власні цілі, ступінь небезпеки та вимагає специфічних підходів до виявлення й нейтралізації.

Аналіз традиційних методів протидії спаму засвідчив, що сигнатурні фільтри та чорні списки залишаються важливим елементом базового захисту завдяки простоті реалізації, високій швидкодії та низьким обчислювальним витратам, але водночас є вразливими до поліморфних і обфускованих атак. Статистичні підходи, зокрема наївний Байєс і TF-IDF, демонструють вищу гнучкість та точність, однак чутливо реагують на зміну мовних патернів, навмисні лексичні викривлення та поєднання кількох мов у межах одного повідомлення. Порівняльна таблиця методів показала, що зі зростанням вимог до точності та адаптивності різко зростає і вартість обробки, що особливо помітно для систем, які працюють у високонавантажених середовищах.

Розгляд методів машинного та глибинного навчання показав, що саме ці підходи сьогодні забезпечують найкраще співвідношення між точністю виявлення спаму та здатністю адаптуватися до нових типів атак. Класичні моделі машинного навчання (SVM, Random Forest та інші) дозволяють виявляти складні нелінійні залежності між ознаками повідомлень і досягати точності, близької до 99%, однак потребують регулярного перенавчання на актуальних даних. Глибинні архітектури – CNN, RNN/LSTM, BERT – дають змогу моделювати контекст, семантику та структуру тексту, автоматично будувати ознаки й показують ще вищі результати, але вимагають потужного апаратного забезпечення, значних обчислювальних ресурсів і складніші в експлуатації та поясненні. У підсумку було обґрунтовано, що найбільш збалансовані результати дають гібридні й ансамблеві системи, які поєднують сигнатурні, статистичні, класичні алгоритми машинного навчання та глибинні моделі в межах єдиної багаторівневої архітектури.

Окремим результатом розділу стало усвідомлення того, що сучасні спам-загрози невіддільні від загального тренду розвитку кіберзлочинності й зростання ролі генеративного штучного інтелекту, ботнетів, evasion та data poisoning атак. Це означає, що протидія спаму виходить за межі суто технічного питання вибору «найкращого алгоритму» й перетворюється на комплексну проблему, де перетинаються технічні, організаційні, правові та етичні аспекти. Теоретичні узагальнення, зроблені у цьому розділі, створюють основу для наступних частин роботи, де буде зосереджено увагу на виборі конкретних моделей машинного та глибинного навчання, побудові експериментальної системи фільтрації спаму та оцінюванні її ефективності в умовах, наближених до реальної експлуатації.

Розділ 2. МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИЯВЛЕННЯ СПАМУ: АЛГОРИТМИ, МОДЕЛІ ТА ПОРІВНЯЛЬНА ОЦІНКА

2.1 Методи класифікації повідомлень у системах протидії спаму

Баєсівські методи класифікації спираються на теорему умовної ймовірності та оцінюють, наскільки ймовірно, що повідомлення є спамом за умови наявності в ньому певних слів. Під час навчання система рахує, як часто кожне слово трапляється у спам-повідомленнях і в легітимній кореспонденції, після чого для слова формується вага як індикатора спаму. Під час класифікації нового листа алгоритм послідовно враховує внесок кожного слова у загальну ймовірність належності повідомлення до класу спаму та порівнює отримане значення з порогом.

Перевага такого підходу полягає у простоті реалізації, лінійній обчислювальній складності та легкому перенавчанні. Оновлення моделі зводиться до корекції статистики появи слів на основі нових прикладів, що дозволяє підтримувати актуальність фільтра без повної перебудови системи. На практиці наївний Байєс стабільно дає точність виявлення спаму на рівні приблизно 95–97% і може обробляти великі обсяги пошти з дуже малими затримками, тому часто використовується як перший, швидкий рівень фільтрації [1].

Ключове обмеження полягає в припущенні незалежності ознак. Алгоритм вважає, що поява кожного слова не залежить від інших слів, тоді як у реальних текстах контекст і сполучуваність мають велике значення. Через це модель гірше працює на складних, контекстно насичених повідомленнях та виявляється вразливою до маніпуляцій на кшталт навмисного додавання типово «безпечних» фраз. Розширені варіанти наївного Байєса (мультиноміальний, Бернуллі) частково зменшують ці недоліки за рахунок урахування частоти та присутності термів, однак принципове припущення про незалежність ознак залишається [24].

Дерева рішень будують модель у вигляді ієрархії послідовних перевірок ознак повідомлення. У корені дерево містить весь навчальний набір, далі алгоритм обирає

ознаку та поріг, які найкраще розділяють спам і легітим за мірою зменшення ентропії або іншого критерію нечистоти. Процес повторюється рекурсивно для кожного вузла, поки підмножини даних не стануть достатньо «чистими» або не буде досягнута гранична глибина [1, 14].

Основна сильна сторона дерев рішень – висока інтерпретованість. Отриману модель можна прочитати як набір зрозумілих правил виду «якщо–то», що спрощує аудит, пояснення рішень та інтеграцію з експертними політиками. При цьому дерева здатні враховувати як бінарні, так і числові та текстові ознаки. У задачах фільтрації спаму правильно побудовані та скорочені дерева зазвичай забезпечують точність на рівні 90–95% [5, 14].

Разом з тим дерева схильні до перепідгонки, особливо якщо дозволити їм рости без обмежень глибини. У такому випадку модель починає запам’ятовувати випадкові деталі навчального набору і втрачає здатність узагальнювати на нові приклади. Для пом’якшення цієї проблеми застосовують процедури обрізання гілок, обмеження глибини, мінімальний розмір вузла та окремі валідаційні набори. Навіть із цими заходами окреме дерево рішень часто поступається за стабільністю більш сучасним методам.

SVM розв’язують задачу класифікації в геометричній постановці. Алгоритм шукає гіперплощину, яка максимально розділяє об’єкти двох класів у просторі ознак, збільшуючи відстань до найближчих точок обох класів. Якщо в початковому просторі класи лінійно не розділювані, застосовуються ядрові перетворення для неявного перенесення даних у простір вищої розмірності, де таке розділення стає можливим [1].

У задачах виявлення спаму SVM демонструють високу точність, наближену до 98–99%, та добре працюють у високовимірних просторах, характерних для текстових даних. Моделі SVM стійкі до частини адаптивних змін у структурі спам-повідомлень, оскільки рішення ґрунтується на загальній геометрії розподілу ознак, а не на окремих ключових словах. Вони також часто краще поводяться на дисбалансованих вибірках, коли частка спаму значно перевищує частку легітимних листів або навпаки.

Основні труднощі при використанні SVM пов’язані з вибором ядра та налаштуванням гіперпараметрів. Невдалий вибір параметрів може призвести як до перепідгонки, так і до недонавчання. Крім того, час навчання різко зростає зі збільшенням кількості навчальних прикладів. Додатковою проблемою є слабка інтерпретованість: пояснити користувачеві, чому конкретне повідомлення класифіковане як спам, зазвичай значно складніше, ніж у випадку дерев рішень або наївного Байєса [16].

Ансамблеві методи класифікації розв’язують задачу іншим підходом, поєднуючи велику кількість відносно простих моделей у єдину систему. Ідея полягає в тому, що сукупне рішення групи різнорідних класифікаторів часто є точнішим і стабільнішим, ніж рішення будь-якої окремої моделі. У контексті фільтрації спаму ансамблі дають змогу одночасно використовувати різні уявлення про дані й компенсувати типові помилки окремих алгоритмів [5, 6].

Random Forest реалізує ансамблеву стратегію через побудову багатьох дерев рішень на різних випадкових підвибірках даних та ознак. Кожне дерево голосує за свій варіант класифікації, а підсумкове рішення приймається за більшістю голосів. Така схема знижує ризик перепідгонки, характерний для одного дерева, та забезпечує високу точність – у задачах виявлення спаму показники можуть досягати 99% і вище при низькій частці помилкових спрацьовувань. Вартість такого підходу – збільшені вимоги до пам’яті та часу класифікації, оскільки повідомлення потрібно послідовно пропустити через десятки або сотні дерев [17].

Бустингові алгоритми, такі як XGBoost, LightGBM чи CatBoost, будують ансамбль послідовно. Кожна нова модель навчається на помилках попередніх, поступово покращуючи якість класифікації на найскладніших прикладах. У задачах фільтрації спаму бустинг зазвичай демонструє точність на рівні 98–99,5% і добре виявляє складні нелінійні залежності. Водночас такі моделі чутливі до налаштування гіперпараметрів і мають вищу обчислювальну вартість навчання, ніж Random Forest.

Стекінг використовує прогнози кількох базових класифікаторів як вхід для окремого метакласифікатора. Така схема дозволяє моделі «навчитися», у яких випадках довіряти тому чи іншому алгоритму, а в яких – коригувати його. Стекінг забезпечує високу гнучкість і дає змогу комбінувати Байєс, дерева, SVM, нейронні мережі в одній системі. Однак за точністю він не завжди суттєво перевищує добре налаштований Random Forest або XGBoost, натомість потребує набагато більших ресурсів на навчання та ускладнює інтерпретацію.

Таким чином, у сучасних антиспам-системах Байєсівські класифікатори залишаються простим і швидким базовим інструментом, дерева рішень забезпечують інтерпретованість, SVM дають високу точність у високовимірних просторах, а ансамблеві методи, насамперед Random Forest та бустинг, задають найкращі показники якості за рахунок комбінування сильних сторін різних підходів. Вибір конкретного методу або їхньої комбінації залежить від вимог до точності, швидкодії, пояснюваності та наявних обчислювальних ресурсів.

2.2 Використання глибоких нейронних мереж у задачах фільтрації спаму

Глибокі нейронні мережі сформували окремий напрям у задачі виявлення спаму, оскільки дозволяють автоматично виділяти інформативні ознаки з тексту без ручного прописування правил та конструювання ознак експертами. Під час навчання такі моделі формують багаторівневі внутрішні представлення, які поступово переходять від простих лексичних патернів до більш абстрактних семантичних структур, що відображають характерні риси спам-повідомлень. Це дає можливість виявляти навіть ті типи спаму, які суттєво відрізняються від прикладів, на яких система навчалася.

CNN застосовуються до тексту за принципом ковзних вікон, коли фіксовані фільтри послідовно проходять по векторних поданнях слів і реагують на локальні фразові конструкції. Для кожного вікна (наприклад, 3–5 слів) обчислюється скалярна

активація, далі застосовується нелінійність ReLU, після чого max-pooling виділяє максимальне значення для кожного фільтра, фактично «фіксуючи» найсильніший збіг із певним патерном у всьому повідомленні. Наявність десятків або сотень фільтрів дозволяє моделі паралельно виявляти різні типи шаблонів, пов’язаних із рекламними формулюваннями, фішинговими зверненнями, типовими тригерами спаму. На стандартних корпусах на кшталт Enron чи SpamAssassin такі архітектури зазвичай показують точність на рівні 95–97% при дуже малій затримці обробки одного листа, що робить їх придатними для високонавантажених поштових сервісів. Обмеження CNN полягає в тому, що вони добре моделюють локальні взаємодії слів, але гірше враховують далекі залежності в довгих текстах, коли важливі фрагменти розташовані далеко один від одного [10, 16].

RNN пропонують іншу стратегію, опрацьовуючи текст послідовно та підтримуючи прихований стан, який акумулює інформацію про попередні слова. На кожному кроці поточне вбудовування слова поєднується з попереднім прихованим станом, формуючи новий стан, що теоретично містить інформацію про всю попередню частину повідомлення. Така конструкція дозволяє відслідковувати довші контексти, але класичні RNN стикаються з проблемою зникання градієнтів: під час зворотного поширення похибка швидко деградує на великих відстанях, через що модель погано навчається залежностям між віддаленими фрагментами тексту. Для задачі фільтрації спаму це означає втрату важливих логічних зв’язків між, наприклад, «обіцянкою виграшу» на початку листа та закликом перейти за посиланням наприкінці.

Архітектура LSTM була запропонована як стабільніший різновид RNN, здатний зберігати релевантну інформацію на сотні кроків уперед. Усередині LSTM-осередку підтримується окремий вектор довготривалої пам’яті, а спеціальні «затвори» регулюють, яку частину попередньої пам’яті зберегти, яку нову інформацію додати та яку частину стану віддати далі. Така побудова забезпечує близьку до одиниці похідну для шляху крізь стан пам’яті, що істотно зменшує проблему зникання градієнтів. На

практиці односпрямовані LSTM-мережі досягають точності близько 96–97,5% на корпусах Lingspam та SPMDC і значно краще відтворюють довгі залежності між контекстом і фінальними фішинговими тригерами.

Двонапрямлені LSTM (BiLSTM) додатково підсилюють можливості моделі за рахунок паралельної обробки послідовності в прямому та зворотному напрямках. Підсумкове представлення кожної позиції поєднує інформацію як про попередні слова, так і про наступні, що особливо важливо у випадках, коли ознаки спаму «оточують» ключову дію з обох боків. У задачах класифікації спаму BiLSTM(двонапрямлена LSTM-архітектура) зазвичай дає приріст точності на 1–2 відсоткові пункти порівняно з односпрямованою LSTM, але вимагає більше пам’яті та часу як на навчання, так і на інференс, оскільки обробка здійснюється вдвічі.

Трансформерні моделі радикально змінили підхід до роботи з послідовностями, замінивши рекурентну структуру на механізм самоуваги. Для кожної позиції генеруються вектори запиту, ключа та значення, на основі яких обчислюються ваги уваги до всіх інших позицій; підсумкове представлення слова є зваженою сумою представлень інших слів із урахуванням цих ваг. Така схема дозволяє моделі безпосередньо «бачити» весь текст одночасно і гнучко формувати залежності між будь-якими позиціями, незалежно від їхньої відстані. Повна паралелізація обчислень робить тренування трансформерів значно ефективнішим на сучасних апаратних платформах, хоча квадратична залежність по довжині послідовності накладає обмеження на максимальний розмір вхідних текстів.

Модель BERT є одним із найуспішніших представників трансформерних архітектур, попередньо натренованим на великих текстових корпусах за допомогою завдань маскованого відновлення слів та прогнозування наступного речення. Завдяки такому навчанні BERT формує контекстно-залежні векторні подання слів, які враховують оточення з обох боків. Для задачі фільтрації спаму поверх BERT зазвичай додають простий класифікаційний шар і проводять донавчання на спеціалізованому датасеті. Експериментальні результати показують, що такі моделі досягають точності

близько 97% при високому F1-значенні й добре узагальнюють на нові типи спаму, навіть якщо спамери навмисно змінюють формулювання або орфографію. Основним недоліком лишаються значні обчислювальні витрати та складність використання на ресурсно обмежених системах [6, 9].

Гібридні архітектури на кшталт комбінації CNN та BiLSTM поєднують переваги різних підходів. CNN-шари спочатку виділяють локальні фразові патерни й стискають інформацію про них у компактні представлення, після чого BiLSTM-шари моделюють послідовні залежності між цими фразами у прямому та зворотному напрямках. На виході формується вектор, який враховує і локальну структуру, і глобальний контекст, що дозволяє досягати точності на рівні 98–99% на різних корпусах спам-повідомлень. Ці рішення є технічно складнішими та вимогливішими до ресурсів, проте дають дуже збалансоване поєднання якості й стійкості до варіацій у тексті [16, 17].

Окремий виклик для всіх описаних моделей створює поява великих мовних моделей, подібних до ChatGPT, Claude, Gemini, які спамери можуть використовувати для генерації масового, граматично коректного та стилістично природного фішингового контенту. Такі листи часто не містять типових орфографічних та стилістичних помилок, за якими раніше легко ідентифікували спам, і за своїми мовними характеристиками наближаються до легітимної ділової кореспонденції. Внаслідок цього навіть моделі рівня BERT, натреновані на схожих корпусах, інколи сприймають подібні повідомлення як «нормальні», що підкреслює необхідність комбінувати глибинні мовні моделі з додатковими шарами перевірки – аналізом поведінкових патернів відправника, перевіркою репутації доменів, виявленням аномалій у мережевій активності тощо.

2.3 Обробка природної мови для виявлення спаму

Обробка природної мови (NLP) у задачах виявлення спаму використовується для перетворення сирого тексту повідомлень на набір формалізованих ознак,

придатних для подальшого машинного аналізу та класифікації. У центрі такого підходу лежить виділення лінгвістичних і статистичних характеристик, які дозволяють надійно відрізняти спам від легітимної кореспонденції [32, 37].

Токенізація є початковим етапом обробки тексту та полягає в розбитті повідомлення на окремі токени – зазвичай слова або стійкі вирази, інколи символи. На цьому кроці алгоритм визначає межі між словами, орієнтуючись на пробіли, пунктуацію та службові символи, після чого часто виконується додаткове очищення: видаляються HTML-теги, зайві пробіли, спеціальні символи, а текст переводиться до нижнього регістру. Така попередня нормалізація зменшує «шум» і спрощує подальші етапи аналізу [8, 27].

Стемінг використовується для механічного скорочення слів до їхніх кореневих форм (стемів) за допомогою набору правил відтинання суфіксів і префіксів. Різні словоформи на кшталт «фільтрування», «фільтрував», «фільтром» зводяться до спільного кореня «фільтр», що зменшує розмірність простору ознак і робить модель менш чутливою до граматичних варіацій. Разом з тим класичні алгоритми стемінгу не враховують частину мови та контекст, можуть породжувати некоректні з погляду мови форми та частково втрачати семантичну інформацію [24, 27].

Лематизація є більш точним і контекстно-залежним способом нормалізації, оскільки повертає слово до його словникової форми з урахуванням морфології та ролі в реченні. На відміну від стемінгу, лематизатор спирається на морфологічний словник і граматичний аналіз, тому результуючі форми завжди є коректними словами мови (наприклад, «є», «був», «буду» → «бути»). Це дозволяє зберігати семантичні відмінності та будувати більш осмислені ознаки, що особливо важливо для багатомовних систем та аналізу складних текстів.

Методи векторизації слів (word embeddings) забезпечують перехід від індексних або «мішкових» подань до щільних векторів у багатовимірному просторі, де близькість між векторами відображає семантичну подібність слів. Моделі Word2Vec (CBOW, Skip-gram) та GloVe навчаються на великих корпусах і розташовують

лексеми так, що слова з подібним вживанням у текстах мають близькі координати. У задачах спам-фільтрації це дозволяє алгоритмам розуміти, що слова «натисніть», «перейдіть», «клацніть» або їхні варіації виконують схожу роль у фішингових сценаріях, навіть якщо їхнє точне написання відрізняється [28].

TF-IDF є класичним статистичним методом оцінки «важливості» слова в документі з урахуванням його частоти в усьому корпусі. Компонента TF відображає, наскільки часто термін трапляється у конкретному повідомленні, тоді як IDF знижує вагу слів, що часто з’являються в більшості документів, і підвищує вагу рідкісних, але показових термів. Для задачі виявлення спаму це означає, що слова, характерні саме для спам-повідомлень і рідкісні в легітимній пошті, отримують вищі ваги й стають ключовими ознаками для класифікації.

Видалення стоп-слів використовується для усунення зайвих за змістом, але надто частотних лексем, які погано допомагають у розрізненні класів. Типові списки включають службові слова («що», «це», «та», «від» тощо), які з’являються майже в кожному тексті й лише додають шум у простір ознак. Водночас сучасні системи часто формують адаптивні списки стоп-слів на основі конкретного корпусу, а для завдань на кшталт фішингу можуть навмисно зберігати окремі службові слова, які входять до типових шаблонів соціальної інженерії.

Семантичний аналіз виходить за межі простого обліку частот та спільного входження слів і спрямований на виявлення значення та інтенцій тексту. Моделі на основі трансформерів, такі як BERT, GPT, Claude та їх модифікації, використовують механізм самоуваги для одночасного врахування взаємозв’язків між усіма словами в повідомленні. Завдяки цьому вони здатні розпізнавати складні фішингові шаблони, де ключові елементи атаки (обіцянка вигоди, створення терміновості, заклик до дії) можуть бути рознесені по різних частинах листа й розділені нейтральним контентом [5, 9].

На практиці побудова NLP-конвеєра для спам-фільтрації включає кілька послідовних етапів. Спочатку виконується очищення тексту й токенізація, далі –

нормалізація (лематизація або стемінг) і видалення стоп-слів, після чого формується числове подання тексту через TF-IDF, embeddings або контекстні вектори трансформерів. Отримані ознаки подаються на вхід класифікаційним моделям – від наївного Байєса, логістичної регресії й SVM до дерев рішень, ансамблевих алгоритмів та глибоких нейронних мереж.

Гібридні підходи поєднують кілька методів обробки природної мови й різні класифікатори в межах однієї системи, що дозволяє підвищити точність і стійкість до варіацій спаму. Наприклад, базовий модуль може використовувати TF-IDF разом із наївним Байєсом для швидкої попередньої фільтрації, тоді як складніші випадки передаються в модуль на основі embeddings і трансформерних моделей. За результатами експериментальних досліджень такі комплексні рішення досягають точності виявлення спаму на рівні 97–99% на стандартних наборах Enron, SpamAssassin, SMS-корпусах тощо, при цьому зберігаючи прийнятний баланс між якістю класифікації й обчислювальними витратами.

2.4 Постановка експерименту та опис набору даних

У межах проведеного дослідження було сформовано експериментальну модель оцінювання методів класифікації спаму, засновану на використанні відкритого корпусу електронних листів SpamAssassin Public Corpus, який містить реальні повідомлення, зібрані з поштових серверів різних доменів. Структура цього набору даних характеризується наявністю окремих текстових файлів, що включають повний вміст листів разом із заголовками, метаданими та основним текстом повідомлення. Загальний обсяг корпусу становить 6047 електронних листів, серед яких 1813 віднесено до категорії спаму, а 4234 — до категорії легітимних повідомлень, що забезпечує достатній рівень варіативності для побудови класифікаційних моделей [1, 8].

Перед здійсненням процедур моделювання було виконано послідовний комплекс попередньої обробки текстових даних, який охоплює нормалізацію структури повідомлень, видалення HTML-тегів, очищення від службових символів, перетворення тексту до нижнього регістру, токенізацію та усунення стоп-слів. Для уніфікації ознак використовувалося TF-IDF-представлення, що дозволило формувати числові вектори з фіксованою структурою. Підготовлений масив даних було розділено на навчальну та тестову вибірки у співвідношенні 80% до 20%, що забезпечує стандартизовані умови для подальшого порівняльного аналізу ефективності алгоритмів.

Перед побудовою моделей було виконано попередню обробку даних:

- видалення HTML-тегів;
- перетворення тексту у нижній регістр;
- очищення від спеціальних символів;
- токенізація;
- видалення стоп-слів;
- лематизація;
- перетворення у векторні представлення (TF-IDF).

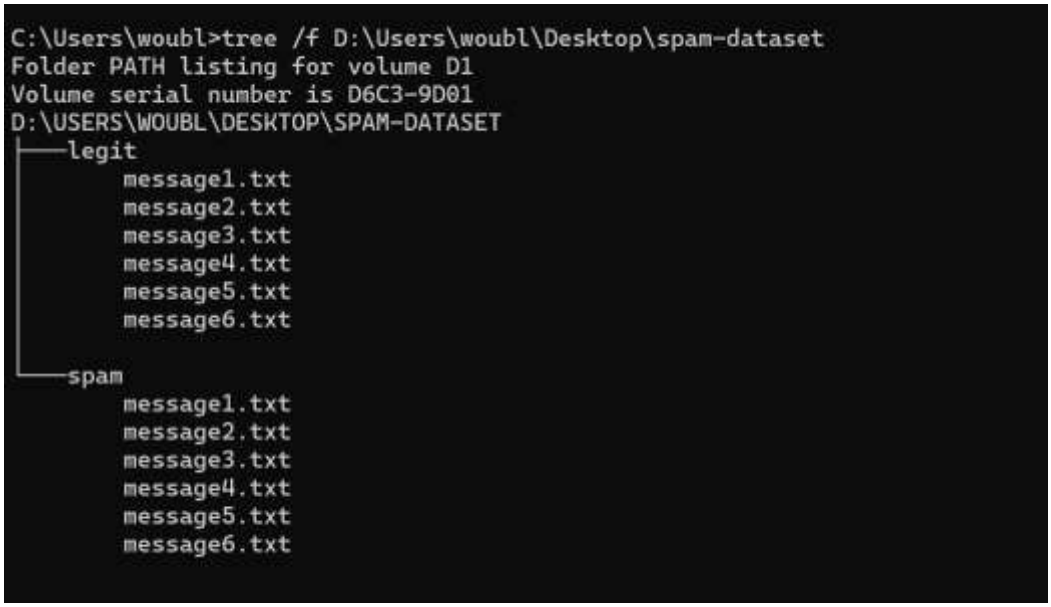

Рисунок 2.1 – Структура набору даних SpamAssassin

Після візуального огляду файлової структури на рис. 2.1 стало можливим встановити характер організації даних, а саме розмежування текстових повідомлень на окремі категорії відповідно до їхнього типу. Такий формат подання забезпечує чітке групування листів і дозволяє безпосередньо контролювати пропорцію спаму та легітимних повідомлень у корпусі. Завдяки цьому зберігається можливість здійснення коректного поділу набору даних на навчальні та тестові підмножини, що є необхідною умовою для формування відтворюваних результатів під час моделювання.

Уміщені в наведеному корпусі файли включають повний текст електронних листів разом із заголовками SMTP, метаданими та основною частиною повідомлень, що дозволяє зберегти як структурні, так і контекстуальні особливості вихідних даних. Таке подання є критичним для задач класифікації, оскільки компоненти електронного листа, зокрема заголовки, поля відправника та ключові фрази, нерідко містять значущі ознаки, які впливають на точність моделі.

Для демонстрації характерних властивостей даних доцільно навести приклад одного з текстових файлів, що входять до корпусу. Представлення окремого повідомлення дозволяє відобразити формат, у якому дані зберігаються до етапу попередньої обробки, і підкреслює необхідність застосування спеціалізованих процедур нормалізації тексту. Приклад вихідного листа подано на рис. 2.2.

```
Шановний співробітник [Компанія 110011],

Наш моніторинг безпеки виявив підозрілу активність у вашому корпоративному акаунті Microsoft 365.
Зафіксовано 3 невдалі спроби входу з невідомих IP-адрес (остання о 14:37 сьогодні) з геолокацій: Польща,
Чехія та VPN-сервери в США. Це може вказувати на компрометацію пароля або атаку brute-force.

Ваш акаунт буде ЗАБЛОКОВАНО через 2 години, якщо не підтвердите свою особу. Це стандартна процедура згідно
з політикою безпеки компанії. Детальніше [link]
```

Рисунок 2.2 – Приклад вихідного листа

Наведений приклад електронного листа демонструє типовий формат подання повідомлень у корпусі та засвідчує наявність значної кількості структурних і контентних елементів, які потребують подальшої нормалізації. Зокрема, у текстових файлах можуть міститися фрагменти HTML-розмітки, спеціальні символи, підписи автоматичних сервісів, некоректно закодовані символи та надлишкові заголовки, які не несуть корисної семантичної інформації для класифікаційної моделі. Такі елементи унеможливлюють безпосереднє використання повідомлень як вхідних даних для алгоритмів машинного навчання, що обумовлює потребу у багатоступеневій процедурі попередньої обробки.

Процес нормалізації тексту є критичним компонентом експериментальної методології, оскільки якість подальших ознак та значення TF-IDF-представлення безпосередньо залежать від коректності очищення та уніфікації текстового матеріалу. Систематизована послідовність етапів попередньої обробки дає змогу забезпечити відтворюваність результатів, зменшити вплив шумових компонентів та вирівняти статистичні властивості текстів для всіх категорій повідомлень.

Рисунок 2.3 – Узагальнена схема застосованого пайплайна попереднього опрацювання

У процесі реалізації експериментальної моделі було розроблено програмний фрагмент, що забезпечує автоматизоване зчитування текстових файлів із локального

корпусу даних, формування відповідних міток класів та перетворення текстових повідомлень у числове представлення. Попередня обробка текстів була реалізована за допомогою бібліотеки *scikit-learn*, яка містить модуль TF-IDF для уніфікованого перетворення текстів у векторні ознаки. Фрагмент програмного коду, який ілюструє процедуру завантаження корпусу та формування TF-IDF матриці ознак, наведено на рис. 2.4.

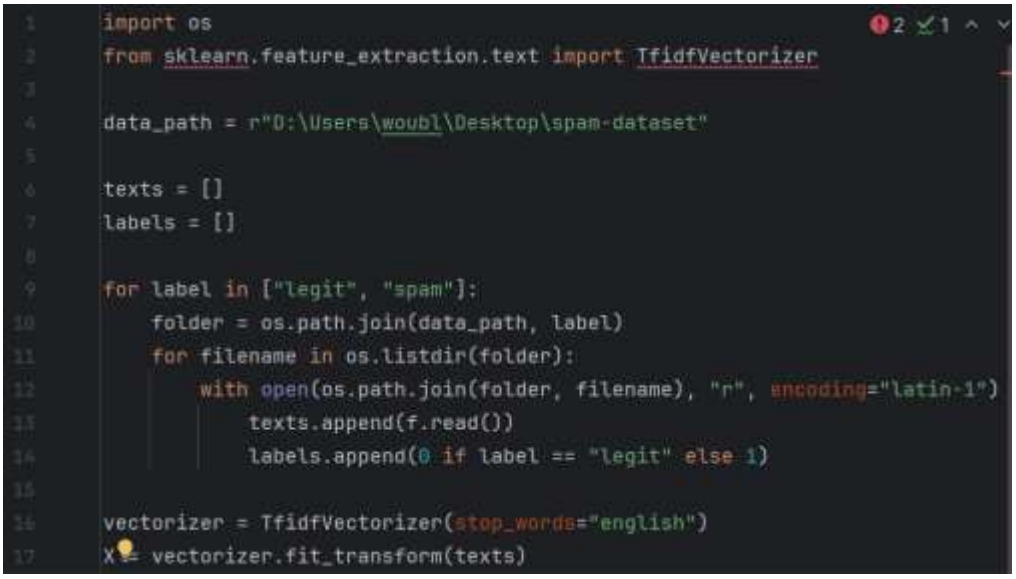

Рисунок 2.4 – Фрагмент коду, що формує TF-IDF матрицю ознак на основі завантаженого корпусу

2.5 Реалізація моделей класифікації спаму

У процесі експериментального оцінювання методів автоматичної класифікації спаму було реалізовано сукупність моделей, що належать до різних підходів машинного та глибинного навчання. Реалізація кожного алгоритму здійснювалася в уніфікованих умовах, які передбачали використання одного й того самого попередньо обробленого корпусу даних, сформованого за допомогою TF-IDF-представлення текстових документів. Єдиним винятком виступала модель рекурентної нейронної мережі LSTM, для якої застосовувалося векторне представлення на основі embedding-

шару. Така стандартизована організація експерименту забезпечила можливість коректного порівняння характеристик моделей за однакових початкових умов.

Першим елементом експериментальної системи було впровадження наївного баєсівського класифікатора мультиноміального типу, що ґрунтується на припущенні умовної незалежності ознак і дозволяє ефективно працювати з текстовими частотними характеристиками. Реалізація моделі здійснювалася за допомогою бібліотеки scikit-learn, що забезпечило її інтеграцію з механізмом TF-IDF-перетворення та надала можливість швидкого формування класифікаційних гіпотез.

Другим алгоритмом, застосованим у рамках дослідження, став метод опорних векторів, налаштований у конфігурації з лінійним ядром. Використання цього підходу дозволило сформувати оптимальну гіперплощину розділення між класами спаму та легітимних повідомлень, що забезпечило високу точність при збереженні стійкості до варіативності текстових структур. Завдяки властивості максимізації маржі SVM проявляє здатність зменшувати кількість хибнопозитивних рішень, що є критично важливим у контексті фільтрації електронної пошти.

Третім компонентом експериментальної моделі став ансамблевий метод Random Forest, який являє собою сукупність незалежних дерев рішень. Кожне дерево формувало прогноз на основі підмножини ознак, що дозволило знизити схильність до перенавчання та забезпечити стабільність результатів при роботі з неоднорідними текстовими даними. Структурна особливість даного методу сприяла підвищенню узагальнюючої здатності класифікатора.

Окреме місце в експерименті посіла рекурентна нейронна мережа типу LSTM, здатна аналізувати послідовні залежності між словами та враховувати контекстні зв’язки, недоступні для класичних статистичних підходів. Модель було побудовано на основі embedding-шару, який формував щільні векторні подання слів, LSTM-компонента, що обробляв послідовність токенів, та вихідного щільного шару з сигмоїдальною функцією активації. Навчання нейронної мережі здійснювалося у

середовищі TensorFlow/Keras із використанням крос-ентропійної функції втрат, що надало можливість отримати високостабільні параметри класифікатора.

Для всіх реалізованих моделей було застосовано однаковий протокол поділу корпусу даних, за яким 80% повідомлень використовувалися для навчання, а решта 20% формували тестову вибірку. Такий підхід забезпечив стандартизовані умови порівняння алгоритмів і дозволив оцінити їх ефективність на основі однорідної множини вхідних ознак.

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nb = MultinomialNB().fit(X_train, y_train)
svm = SVC(kernel="linear").fit(X_train, y_train)
rf = RandomForestClassifier(n_estimators=200).fit(X_train, y_train)
```

Рисунок 2.4 – Фрагмент програмного коду, що демонструє процес навчання моделей класифікації

2.6 Порівняльний аналіз результатів застосування традиційних і сучасних методів

Порівняльний аналіз традиційних та сучасних методів виявлення спаму демонструє, що перехід від класичних алгоритмів машинного навчання до глибинних архітектур є не плавним удосконаленням, а якісною зміною підходу до побудови систем захисту електронної пошти. Дослідження показують, що використання автоматичного вилучення ознак у глибоких моделях дозволяє суттєво підвищити точність, зменшити кількість помилкових спрацьовувань та підвищити здатність системи адаптуватися до нових типів спаму.

Традиційні методи – Наївний Байєс, K-найближчих сусідів, SVM, Random Forest – забезпечують прийнятну якість класифікації на стандартних наборах даних типу Enron і SMS Spam Collection, досягаючи приблизно 85–90% точності для Наївного

Байєса та K-NN, 90–93% для SVM і 92–95% для Random Forest. Їх головним обмеженням є залежність від ручної інженерії ознак: якість моделі суттєво визначається тим, наскільки повно аналітики змогли відібрати й закодувати релевантні текстові характеристики, що ускладнює виявлення складних, контекстно залежних патернів.

Глибокі нейронні мережі, зокрема CNN та LSTM, демонструють помітно вищу точність на тих самих наборах даних завдяки автоматичному вилученню ознак і здатності моделювати складні залежності у тексті. Для CNN, поєднаних з GloVe-вбудовуваннями, досягається точність близько 96–96,5%, тоді як LSTM-архітектури показують результати на рівні 96–97,5% і вище. При об'єднанні кількох відкритих корпусів (Enron, SpamAssassin, SMS Spam Collection) точність глибоких методів зростає приблизно до 96,5–97,5%, тоді як традиційні алгоритми залишаються в межах 90–92%, що відповідає різниці на 10–14 відсоткових пунктів.

Важливою з практичної точки зору є різниця у рівні помилкових позитивних спрацьовувань, коли легітимні повідомлення помилково потрапляють до спаму. Для Наївного Байєса, K-NN та SVM цей показник зазвичай становить 3–5%, що означає втрату від одного до п'яти коректних листів зі ста, тоді як для ансамблевих методів на кшталт Random Forest він знижується до 1,5–2,5%. Глибокі моделі, зокрема CNN з GloVe та LSTM, дозволяють опускати рівень помилкових позитивів до 1–2%, а при використанні добре налаштованих моделей рівня BERT і гібридних систем цей показник може бути меншим за 1%, що є критично важливим для великих поштових сервісів і корпоративних систем.

Часові та обчислювальні витрати формують ще один важливий критерій вибору. Традиційні методи, такі як Наївний Байєс і K-NN, навчаються протягом хвилин навіть на вибірках із сотень тисяч або мільйонів повідомлень, а SVM і Random Forest потребують годин, але залишаються прийнятними для періодичного перенавчання. Глибокі нейронні мережі (CNN, LSTM, BERT) вимагають значно більших ресурсів: навчання може тривати від кількох годин до днів і тижнів на GPU, однак це одноразові

або рідкісні витрати, тоді як класифікація окремих повідомлень у вже натренованих моделях займає мілісекунди, що співставно зі швидкістю традиційних методів.

Гібридні архітектури, які поєднують швидкі традиційні класифікатори для попереднього відсікання очевидного спаму з глибокими моделями для аналізу складних або сумнівних випадків, демонструють найкращі практичні результати. Типова схема передбачає перший рівень на базі TF-IDF + Наївний Байєс або Random Forest і другий рівень з використанням BERT чи CNN+LSTM для поглибленого семантичного аналізу. Такі системи забезпечують загальну точність на рівні 98–99% із часткою помилкових позитивів нижче 1%, залишаючись при цьому придатними до розгортання в реальних високонавантажених середовищах.

Таблиця 2.1

Порівняльний аналіз алгоритмів класифікації спаму за показниками ефективності

Метод	**Точність (%)**	**Помилкові позитиви (%)**	**Час тренування**	**Практична масштабованість**	**Адаптивність**
Наївний Баєс	85–90	3–5	Хвилини	Висока	Низька
K-NN	88–92	2.5–4	Хвилини–години	Середня	Середня
SVM	90–93	2–3.5	Години–дні	Середня	Середня

Random Forest	92–95	1.5–2.5	Години	Висока	Висока
LSTM	96–97.5	1–2	Дні	Середня	Дуже висока
CNN	95–97	1–1.5	Дні	Середня	Висока
BERT	97–98	<1	Години (transfer learning)	Висока	Дуже висока
Гібридні системи	98–99	<1	Комбіновано	Висока	Дуже висока

З практичних позицій традиційні методи доцільно застосовувати в організаціях із обмеженими ресурсами та невеликим обсягом поштового трафіку, де критичним є простий супровід і мінімальні вимоги до обчислювальної інфраструктури. Великі провайдери та корпоративні системи, які обробляють мільйони повідомлень на добу та стикаються з еволюціонуючими загрозами, отримують найбільшу вигоду від гібридних і глибоких рішень, які краще адаптуються до нових типів спаму, включно з контентом, згенерованим великими мовними моделями. У системах, де пропуск одного успішного фішингового листа може призвести до значних фінансових або репутаційних втрат, додаткові витрати на глибинні моделі є виправданими.

Висновки до другого розділу

У другому розділі було проведено комплексний аналіз методів інтелектуального опрацювання даних, що застосовуються для виявлення спаму, а також здійснено порівняльну оцінку ефективності алгоритмів машинного й глибинного навчання у цій сфері. Розгляд класичних моделей — наївного Байєса, дерев рішень, машин опорних векторів та ансамблевих підходів — показав, що попри значні відмінності в архітектурі та принципах роботи, усі ці алгоритми здатні забезпечувати високі показники точності, придатні для практичного використання в системах електронної пошти. Найкращі результати у межах традиційних методів демонструють ансамблеві моделі, зокрема Random Forest та бустингові алгоритми, які завдяки поєднанню різних підходів забезпечують баланс між точністю, стійкістю та здатністю до узагальнення.

Окрему увагу у розділі приділено глибинним нейронним мережам, що сьогодні становлять основу найсучасніших антиспам-рішень. CNN, RNN/LSTM та трансформерні моделі на кшталт BERT забезпечують значно глибше розуміння структури та семантики повідомлень, дозволяючи ефективно виявляти як традиційні, так і адаптивні форми спаму. Показано, що трансформери демонструють найвищу точність і найкращу здатність до узагальнення, хоча потребують істотно більших обчислювальних ресурсів. Гібридні архітектури, що поєднують CNN та BiLSTM, забезпечують оптимальне співвідношення швидкодії й глибини аналізу та становлять перспективний напрям подальших досліджень.

Також у розділі проаналізовано ключові NLP-процедури, які є невід'ємною частиною сучасних систем фільтрації: токенізацію, нормалізацію, вилучення ознак, побудову векторних подань та методи попередньої обробки тексту. Показано, що якість таких процедур напряму визначає ефективність подальших моделей, а отже, є критично важливою складовою успішного проєктування антиспам-систем.

Окремо було приділено увагу практичним аспектам підготовки даних та навчання моделей. Дослідження показало, що використання якісних збалансованих корпусів, застосування технік боротьби з дисбалансом класів, правильний вибір

гіперпараметрів та регулярне оновлення моделей суттєво впливають на підсумкові показники точності, повноти та F1-міри.

У підсумку встановлено, що сучасна протидія спаму неможлива без поєднання класичних методів, алгоритмів машинного навчання, глибинних нейронних мереж та NLP-технологій. Кожен із підходів має власні сильні сторони, а найвищої ефективності можна досягти шляхом використання ансамблевих або гібридних систем, що інтегрують декілька моделей у єдиний механізм фільтрації. Результати, отримані в межах цього розділу, слугують практичною та методологічною основою для побудови експериментальної антиспам-системи та проведення порівняльної оцінки її роботи у третьому розділі.

РОЗДІЛ 3. ІННОВАЦІЙНІ ПІДХОДИ ТА ПЕРСПЕКТИВНІ НАПРЯМИ РОЗВИТКУ АНТИСПАМ-ТЕХНОЛОГІЙ

3.1 Використання квантових технологій у протидії спаму

Використання квантових технологій у протидії спаму розглядається як перспективний напрям, що поєднує досягнення квантової інформатики, квантової криптографії та квантово-прискореного машинного навчання для підвищення точності фільтрації й стійкості інфраструктури електронних комунікацій до атак з боку спамерів. У сучасному науковому дискурсі формується розуміння того, що квантові обчислювальні моделі здатні не лише радикально змінити баланс сил у криптоаналізі, але й створити новий клас інструментів для детектування складних і адаптивних спам-кампаній, які обходять класичні статистичні та нейромережеві фільтри.

Теоретичну основу квантових підходів становить формалізм квантової інформації, де елементарною одиницею виступає кубіт, що може перебувати в суперпозиції станів і утворювати заплутані багатовимірні системи, здатні репрезентувати та паралельно обробляти високорозмірні вектори ознак повідомлень електронної пошти. У межах цієї парадигми класичні моделі навчання, зокрема лінійні класифікатори, метод опорних векторів та глибинні мережі, отримують квантові аналоги, у яких геометричні операції над векторами ознак реалізуються через унітарні перетворення й вимірювання у відповідних базисах, що теоретично забезпечує експоненціальне прискорення окремих підзадач обчислення скалярних добутків, пошуку найближчих сусідів і оптимізації параметрів [29, 30].

Структурний принцип квантового підходу до класифікації спаму може бути представлений у вигляді послідовності взаємопов’язаних етапів, де класична обробка тексту інтегрується зі спеціалізованими квантовими перетвореннями. Типова схема включає етап підготовки даних у вигляді векторних подань, кодування цих векторів у

квантові стани шляхом амплітудного або кутового кодування, застосування унітарних операцій для трансформації простору ознак, виконання квантових обчислень, що відповідають класифікаційним перетворенням, та завершальне вимірювання, яке повертає класову гіпотезу. Такі схеми наведено на рис. 3.1, де відображено загальну логіку взаємодії класичних та квантових компонентів у гібридних антиспам-системах.

Рисунок 3.1 – Узагальнена схема квантового еволюційного алгоритму

У сучасних дослідженнях зі спам-фільтрації демонструються гібридні архітектури, де квантові моделі інтегруються як прискорювачі окремих компонентів класичного конвеєра обробки електронної пошти, насамперед на етапах векторизації, зменшення розмірності та побудови класифікаційних гіперплощин у просторах великої розмірності. Порівняльний аналіз показує, що для високорозмірних та розріджених представлень контенту електронних листів квантові варіанти моделей можуть досягати кращого співвідношення між точністю виявлення спаму та

обчислювальними витратами на великих вибірках, особливо в умовах класового дисбалансу й наявності складних шаблонів обходу традиційних фільтрів.

Окремий напрям розвитку пов’язаний з використанням квантового машинного навчання в контексті кібербезпеки, де квантові моделі розглядаються як інструмент для виявлення аномалій у трафіку, розпізнавання фішингових кампаній та побудови більш стійких профілів поведінки користувачів, що враховують приховані кореляції у великих масивах даних. У застосуванні до протидії спаму це відкриває можливість побудови багаторівневих систем, які одночасно аналізують текстове наповнення, метадані, часові патерни відправлення та мережеві зв’язки відправників, використовуючи квантові схеми для пришвидшення кластеризації та класифікації у багатовимірному просторі ознак.

Попри значний потенціал квантових методів, їх впровадження супроводжується низкою фундаментальних обмежень. До найважливіших належать обмеженість сучасної квантової апаратної бази, висока чутливість кубітів до декогеренції, наявність шумів, складність масштабування квантових схем, а також відсутність промислових квантових процесорів, здатних працювати зі складними високорозмірними задачами класифікації в режимі реального часу. Також суттєвим обмеженням є невизначеність щодо стандартів інтеграції квантових прискорювачів у реальні поштові інфраструктури, оскільки більшість доступних сьогодні квантових прототипів працюють у модельованих або лабораторних умовах і демонструють обмежену стабільність під час багаторазових повторюваних обчислень.

Додатковим фактором ризику є загроза з боку повноцінних квантових комп’ютерів для сучасних криптографічних протоколів, що забезпечують автентифікацію та цілісність електронної пошти. Потенційна можливість зламу алгоритмів з відкритим ключем спричиняє необхідність переходу до постквантових криптографічних стандартів, оскільки ефективність квантових моделей машинного навчання в антиспам-системах залежить від стійкості механізмів автентифікації [30].

Перспективи розвитку квантових антиспам-технологій пов’язуються з появою більш масштабованих квантових процесорів, здатних працювати з великими векторами ознак, з удосконаленням методів кодування класичних даних у квантові стани, зі зростанням точності та стабільності унітарних перетворень, а також із формуванням протоколів безпечної взаємодії між класичними та квантовими підсистемами [29, 30]. У середньостроковій перспективі очікується подальший розвиток гібридних квантово-класичних архітектур, які можуть стати основою для прискореної класифікації та виявлення спаму в системах масової електронної пошти, тоді як у довгостроковому горизонті можливим є перехід до повністю квантових обчислювальних моделей, що здійснюватимуть паралельний аналіз тексту, метаданих і поведінкових патернів у рамках єдиної квантової платформи.

3.2 Блокчейн-рішення в боротьбі зі спамом

Використання блокчейн-технологій у сфері протидії спаму розглядається як системний підхід до формування децентралізованого, стійкого до маніпуляцій середовища, у якому процеси збирання, маркування та валідації спам-контенту відбуваються під повним контролем прозорого реєстру транзакцій. У межах такої моделі блокчейн не виступає самостійним інструментом класифікації, а виконує функцію інфраструктурного рівня, що забезпечує незмінність історії взаємодій, перевірюваність внесених даних та неможливість несанкціонованого коригування результатів моделювання під час навчання або перевірки класифікаторів. Основна логіка організації децентралізованої антиспам-системи на блокчейні представлена на рис. 3.2, де схематично відображено процеси взаємодії між користувачами, смарт-контрактами, оракулами та механізмами зовнішнього машинного навчання [35,36].

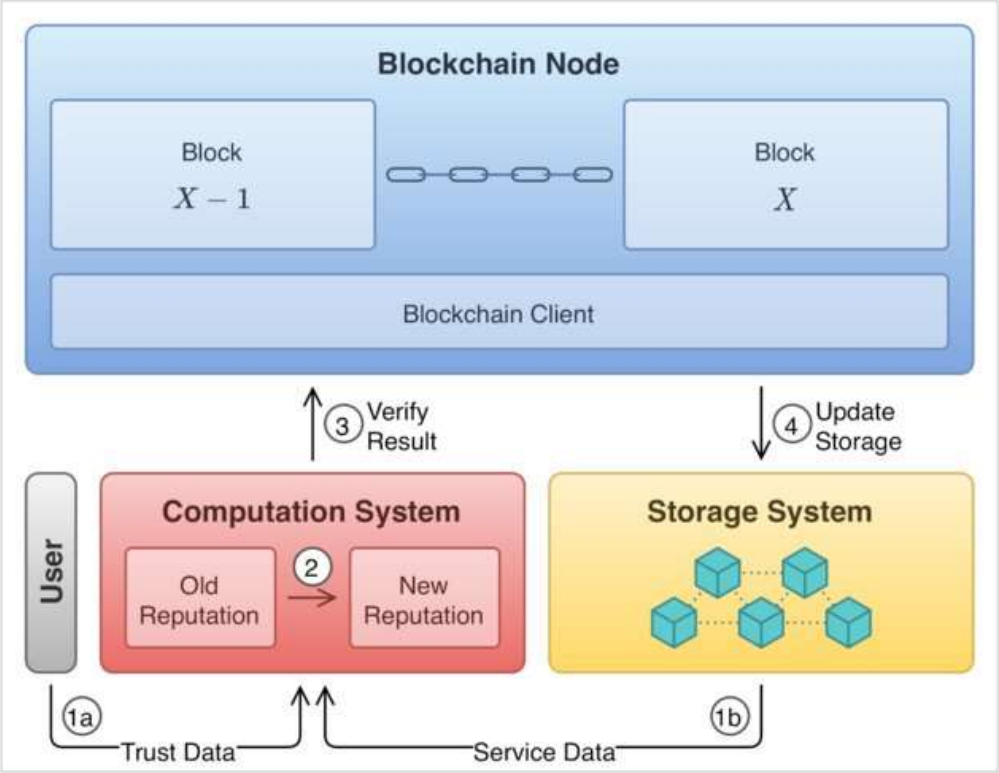

Рисунок 3.2 – Схема взаємодії компонентів децентралізованої антиспам-системи на блокчейні

У зазначеній архітектурі центральну роль відіграють смарт-контракти, які забезпечують децентралізоване керування процесами краудсорсингового маркування, перевірки даних та економічного стимулювання учасників. Внесення нових прикладів спам-контенту супроводжується формуванням транзакції, що фіксує джерело, хеш-ідентифікатор зразка та відповідний стейкінг. Модель машинного навчання, розміщена поза ланцюгом, здійснює оцінювання якості надісланих даних, а результати перевірки повертаються у блокчейн через оракул [23, 31]. Смарт-контракт на основі отриманої інформації автоматично визначає винагороду або штраф, враховуючи внесок окремого учасника в покращення або погіршення точності класифікаційної моделі [22, 31].

У такій системі блокчейн забезпечує прозорість розподілу репутаційних ваг, які використовуються для агрегування рішень великої кількості користувачів. Репутаційні значення, що зберігаються в реєстрі, дозволяють зменшувати вплив

аномальних або зловмисних учасників, тоді як механізми динамічного зважування зразків пригнічують можливість повторного внесення однакових даних або систематичного використання раніше винагороджених прикладів. Це сприяє формуванню стійкої економічної моделі, у якій цінність мають лише унікальні та якісно розмічені повідомлення.

Другим напрямом застосування блокчейну є побудова мікроплатіжних систем, у яких кожне електронне повідомлення супроводжується невеликою транзакцією, що створює фінансовий бар’єр для масових розсилок. У разі класифікації листа як спаму умовна сума не повертається відправнику, що робить великі спам-кампанії економічно нерентабельними. Розширення таких моделей на основі децентралізованих репутаційних механізмів дозволяє автоматично виявляти раніше позначені джерела та формувати додатковий рівень фільтрації шляхом аналізу історії транзакцій конкретних гаманців і доменів.

У контексті веб-спаму блокчейн надає можливість зафіксувати результати модерації або голосування користувачів у незмінному вигляді, що унеможливлює ретроспективне редагування або маніпулювання оцінками сайтів. У такій моделі кожна позначка або висновок моделі зберігається як транзакція, а зв’язок з відповідним ресурсом здійснюється через криптографічно захищений хеш. Це підвищує стійкість до координаційних атак, знижує ризик внутрішніх порушень і надає можливість точно відстежувати історію змін статусу веб-сторінок.

Разом із тим практична імплементація таких систем стикається з низкою структурних обмежень, серед яких обмежена пропускна здатність блокчейнів першого рівня, висока вартість транзакцій, значні часові затримки підтвердження, а також складність забезпечення конфіденційності маркованих даних без порушення принципу прозорості. Ці фактори вимагають застосування гібридних рішень, що передбачають винесення обчислювально затратних операцій за межі ланцюга, використання оракулів, а також інтеграцію платформ другого рівня для оптимізації вартості та швидкості транзакцій.

Перспективи розвитку блокчейн-орієнтованих антиспам-систем пов’язуються з подальшим удосконаленням стимулюючих механізмів, впровадженням адаптивних економічних моделей, здатних реагувати на зміну тактик спамерів, а також з поглибленою інтеграцією децентралізованих підходів із сучасними алгоритмами машинного навчання [6, 31]. Очікується, що розвиток масштабованих блокчейн-платформ, вдосконалення технологій оракулів, а також використання приватних або консорціумних ланцюгів дозволить створити промислові антиспам-системи, у яких прозорість, незмінність і перевірюваність блокчейну поєднуватимуться зі швидкістю й гнучкістю класичних ML-моделей.

3.3 Гібридні системи та інтеграційні рішення

Гібридні антиспам-системи виникають як відповідь на обмеженість окремих методів і передбачають поєднання статистичних, сигнатурних, евристичних і машинно-навчальних технік у багаторівневу архітектуру, здатну реагувати на розмаїття сучасних спам-патернів. Така архітектура забезпечує баланс між швидкодією, точністю та стійкістю до еволюційних змін у тактиках зловмисників, створюючи єдине середовище для аналізу текстових, структурних, мережевих та поведінкових характеристик. Сама концепція гібридності передбачає не просто одночасне використання різних фільтрів, а побудову взаємопов’язаних логічних рівнів, кожен з яких компенсує недоліки інших, формуючи більш комплексну модель взаємодії з потоками вхідних повідомлень. У цьому сенсі гібридна система не прагне замінити окремі модулі, а створює над ними узгоджений механізм інтерпретації та прийняття рішень.

Особливої ефективності гібридні рішення досягають у разі інтеграції кількох незалежних каналів обробки, де кожен модуль спеціалізується на окремій групі ознак: контент-наповнення, репутація відправника, часові особливості, перевірка автентичності доменів за SPF, DKIM і DMARC, дані телеметрії з поштових та мережевих шлюзів. Така багатоканальність дозволяє не лише формувати

багатовимірний простір ознак, але й виділяти латентні залежності, які невидимі окремим моделям. Крім того, багатоканальна структура забезпечує можливість ізоляції помилкових сигналів, коли хибнопозитивні або хибнонегативні результати одного з модулів нейтралізуються інформацією, отриманою з інших джерел. Це створює передумови для суттєвого підвищення загальної стійкості та зниження ризику пропуску спаму в умовах високої варіативності загроз [6].

На цьому етапі доречно представити авторську модель, яка узагальнює принципи гібридної фільтрації та демонструє взаємодію модулів у динамічній системі.

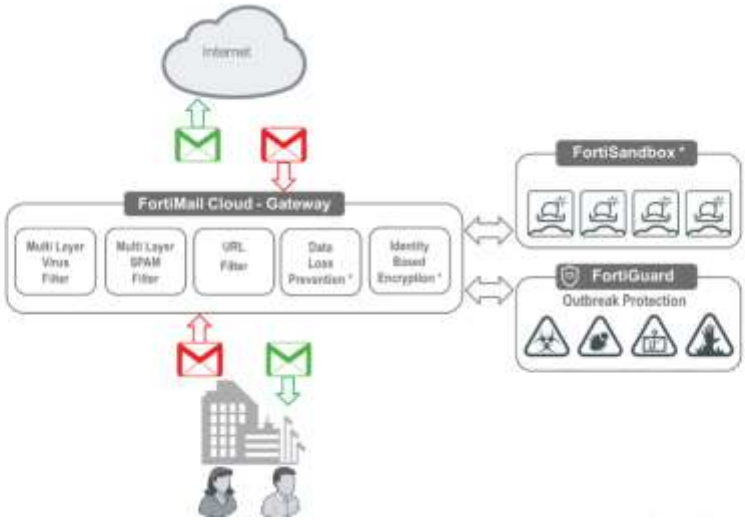

Рисунок 3.3 – Узагальнена схема авторської моделі гібридної системи фільтрації електронної пошти

Після формалізації архітектурної моделі стає можливим розглянути її адаптивні властивості. У контексті практичного функціонування адаптивність відіграє ключову роль, оскільки саме вона забезпечує здатність моделі реагувати на динамічні зміни у структурі трафіку, інтенсивності спам-розсилок та появі нових методів обходу фільтрів. У типових умовах система працює у базовому режимі, де пріоритет віддається легким сигнатурним і евристичним перевіркам, оптимізованим для низьких затримок. Однак у разі зміни профілю ризику система переходить у розширений режим. Це стосується ситуацій, коли спостерігається різке збільшення

семантично споріднених повідомлень, поява вкладень нетипових форматів, формування нових ланцюжків редиректів або відхилення від усталеної поведінки відправника чи домену.

У таких випадках активується набір глибинних моделей, які аналізують текстове наповнення з урахуванням контексту та прихованих залежностей між фразами. Паралельно робота корелюється з даними SIEM-платформи, що дозволяє виявляти багатокрокові сценарії атак, коли спам є лише першим, але критично важливим елементом складнішого ланцюга. Подібні сценарії характерні для поширення шкідливого ПЗ, фінансового шахрайства або соціальної інженерії, що часто використовує електронну пошту як первинний канал комунікації.

Цілком очевидно, що така система потребує чітко організованого механізму взаємодії між рівнями. Саме тому у сучасних рішеннях дедалі частіше застосовуються підходи модульної архітектури, коли кожен компонент має чітко визначений інтерфейс і обмежену відповідальність. Це забезпечує можливість незалежного оновлення одного або кількох модулів без необхідності переробляти весь конвеєр. У межах такої архітектури результати кожного модуля нормалізуються до уніфікованого показника довіри, що надалі обробляється метамоделлю. Таким чином досягається не лише узгодженість, але й покращена інтерпретованість поведінки системи, що є суттєвим аспектом для випадків, коли необхідно аналізувати причини хибних спрацьовувань.

Розвиток гібридних систем також пов’язаний із поступовим переходом до розподілених моделей обробки, коли частина функцій виконується на рівні клієнтських вузлів, а частина — у хмарному середовищі. Це дозволяє ефективно балансувати обчислювальні ресурси, зменшувати затримки та уникати перевантаження центральних серверів. Особливої актуальності такі рішення набувають у корпоративних середовищах із великою кількістю користувачів і географічно розподіленими робочими станціями.

У цілому гібридні та адаптивні системи фільтрації формують наступне покоління антиспам-технологій, поєднуючи різноманітні методи аналізу, гнучку багаторівневу логіку та високий ступінь автоматизації, що дозволяє оперативно реагувати на постійні зміни у структурі загроз та мінімізувати ризики пропуску шкідливих або небажаних повідомлень.

3.4 Концептуальна модель інтелектуальної системи протидії спаму

Концептуальна модель інтелектуальної системи протидії спаму розглядається як узагальнена архітектура, що охоплює сукупність методів, підходів і механізмів, здатних забезпечити високий рівень точності в умовах постійної еволюції спам-технологій. В основі цієї моделі лежить ідея багаторівневої організації, у межах якої різні категорії алгоритмів виконують взаємодоповнювальні функції, формуючи цілісну структуру інтелектуального аналізу електронних повідомлень. Така побудова дозволяє не лише компенсувати недоліки окремих моделей, але й створити середовище, орієнтоване на активне накопичення, переосмислення та інтеграцію знань, що відображають динаміку загроз у реальному часі.

На базовому рівні модель передбачає всебічну нормалізацію й уніфікацію вхідних даних, що включає очищення повідомлень від шумів, стандартизацію структури листів, вилучення текстових і нетекстових компонентів, а також первинний аналіз метаданих, отриманих із джерел автентифікації. Цей рівень відіграє критичну роль у формуванні єдиної інформаційної основи для подальшої обробки, оскільки навіть незначні відхилення у форматі повідомлень можуть призвести до некоректного трактування їх вмісту на рівні моделей машинного навчання. Крім того, первинна уніфікація дає можливість виділити ключові патерни на ранніх етапах і зменшити обсяг обчислень, необхідних для глибинного аналізу [1].

Другий рівень становить ядро інтелектуальної обробки, яке організоване як гетерогенна сукупність моделей, кожна з яких спеціалізується на власному типі ознак.

Статистичні моделі забезпечують швидку оцінку на основі частотних характеристик тексту; ансамблеві методи дозволяють структурувати інформацію у високорозмірних просторах ознак; глибинні нейронні мережі інтерпретують семантичні залежності; трансформерні архітектури працюють із контекстом повідомлення в його глобальному й локальному вимірах; поведінкові й мережеві моделі аналізують властивості трафіку та взаємодій у просторі доменів. У такій конфігурації кожен елемент системи генерує власну оцінку ризику, а конкуренція між цими оцінками забезпечує збалансований і стійкіший результат порівняно з використанням окремої моделі [8].

Надбудованим елементом виступає інтегративний рівень, відповідальний за прийняття кінцевого рішення. Він поєднує результати різнорідних моделей у єдину шкалу довіри, ураховуючи історію попередніх взаємодій із конкретним відправником, метадані, що характеризують зміну поведінки домену, глобальні сигнали репутації, а також локальні ознаки, отримані від попередніх рівнів. Оскільки структура багаторівневого трафіку передбачає наявність складних залежностей між подіями, цей рівень реалізує логіку контекстної інтерпретації, у якій повідомлення оцінюється не ізольовано, а в межах ширшого інформаційного сценарію. Таким чином досягається значне підвищення точності щодо фішингових кампаній, які часто маскуються під легітимні повідомлення, але розкривають свою шкідливу природу лише у взаємозв’язку з попередніми діями зловмисника [17].

Наступний рівень, а саме рівень самоадаптації — забезпечує безперервне вдосконалення системи завдяки використанню механізмів активного та напівавтоматичного навчання. Інтелектуальна система здатна визначати повідомлення з високим рівнем невизначеності та передавати їх на ручну верифікацію, а отримані користувачем мітки повертаються в цикл навчання моделей. Система враховує хибнопозитивні та хибнонегативні рішення, коригує внутрішні параметри, переглядає ваги моделей і перебудовує структуру пріоритетів залежно від змін у структурі спаму. Таким чином формується динамічний механізм еволюційної

адаптації, здатний реагувати на нові спам-тактики незалежно від їхньої структури та інтенсивності.

У межах концептуальної моделі особливе місце займає компонент пояснюваності рішень, необхідний для забезпечення прозорості роботи системи та можливості її аудиту. З огляду на складність глибинних моделей і трансформерних архітектур, прозорість рішень є критично важливою для розуміння того, які саме ознаки або фактори вплинули на класифікацію повідомлення. У цій моделі пояснюваність реалізується через побудову багаторівневих карт впливу, оцінку значущості ознак та аналіз проміжних представлень, що дозволяє операторам системи виявляти потенційні помилки в роботі моделей та вдосконалювати їх у майбутньому.

Перспективною складовою концептуальної моделі є її здатність до інтеграції з зовнішніми інфраструктурами, такими як блокчейн-мережі та квантово-орієнтовані обчислювальні середовища. Використання блокчейну дозволяє фіксувати репутаційні сигнали, гарантувати незмінність історії взаємодій та формувати децентралізований механізм зберігання даних, що підвищує стійкість системи до спроб маніпуляцій. Інтеграція з квантовими обчисленнями, хоча наразі залишається перспективною, потенційно відкриває можливості для суттєвого пришвидшення окремих підзадач — зокрема пошуку оптимальних представлень та кластеризації у високорозмірних просторах ознак [29].

Узагальнюючи викладене, концептуальна модель інтелектуальної системи протидії спаму являє собою багаторівневу, гнучку та здатну до самоадаптації структуру, у якій синтезуються різні парадигми аналізу даних, методи глибинного навчання, інструменти моделювання поведінки та зовнішні механізми забезпечення довіри. Її архітектура відображає сучасні тенденції розвитку кіберзахисту та відповідає потребам інформаційної інфраструктури, що функціонує в умовах високої варіативності загроз. Така модель має потенціал стати основою для створення систем протидії спаму нового покоління, здатних не лише виявляти очевидні та масові атаки, але й ідентифікувати приховані та адаптивні схеми, що постійно видозмінюються.

3.5 Майбутні виклики та напрями досліджень

Перспективний розвиток антиспам-технологій визначається якісною зміною загрозового ландшафту, насамперед через масове використання генеративних моделей штучного інтелекту для створення фішингових і спам-кампаній, що відрізняються високим ступенем персоналізації, граматичною бездоганністю та здатністю до масштабування. Сучасні спостереження фіксують стрімке зростання обсягів атак, пов’язаних із використанням генеративних мовних моделей, при цьому якість таких повідомлень зрівнюється з контентом, підготовленим людиною, що нівелює традиційні евристики, орієнтовані на грубі стилістичні та лінгвістичні помилки.

Подальші дослідження у цій сфері неминуче зосереджуються на розробленні семантично чутливих механізмів детектування, здатних аналізувати намір і контекст повідомлень, а не лише поверхневі текстові чи технічні ознаки. На перший план виходять моделі, що поєднують глибинні трансформерні репрезентації з аномалійним аналізом поведінки користувачів і комунікаційних патернів, а також підходи до інтерпретованості, що дозволяють пояснювати рішення фільтрів у термінах семантичних відхилень і нетипових комунікаційних сценаріїв.

Окремий блок викликів пов’язаний із еволюцією атак ухилення (evasion attacks), спрямованих на цілеспрямоване конструювання ворожих прикладів, які залишаються зрозумілими для людини, але систематично вводять в оману моделі класифікації спаму. Для текстових і NLP-орієнтованих фільтрів демонструється, що незначні модифікації на рівні символів, слів або речень — включно з варіантами типу «sp@m» — здатні різко знижувати точність як класичних алгоритмів, так і глибинних мереж [20].

У контексті антиспаму це означає, що розвиток моделей детектування має невід’ємно супроводжуватися побудовою та стандартизацією процедур тестування на стійкість до ворожих прикладів, а також створенням спеціалізованих наборів даних із включенням реалістичних прикладів ухилення. Перспективним напрямом є

поєднання тренування з урахуванням ворожих прикладів, використання ансамблів моделей з різною чутливістю до типів змін та застосування багаторівневих механізмів «sanity-check» для виявлення аномально сконструйованих текстів і URL-адрес.

Не менш важливою проблемою постає баланс між точністю, продуктивністю та інтерпретованістю складних глибинних архітектур, що домінують у сучасних системах виявлення фішингу та спаму. Огляд новітніх моделей виявляє тенденцію до зростання обчислювальної вартості, потреб у даних та складності внутрішніх репрезентацій, що ускладнює їх розгортання в ресурсно обмежених середовищах та інтеграцію в критичні бізнес-процеси. У відповідь формується запит на більш «обмежувані» архітектури з вбудованими засобами пояснення рішень, використанням методів пояснюваного ШІ та змішаних підходів, де високоточні, але «важкі» моделі працюють у зв’язці з легковаговими попередніми фільтрами [19].

Синергія між антиспам-технологіями та системами корпоративного захисту інфраструктури також створює нові дослідницькі завдання, пов’язані з побудовою мультиканальних і мультимодальних платформ протидії соціальній інженерії, що охоплюють електронну пошту, месенджери, голосові та відеосервіси. Інтеграція даних про текст, голос, зображення, а також поведінкові сигнали користувачів в єдині моделі детектування вимагає нових методів об’єднання ознак, спільного навчання та узгодженої оцінки ризиків у реальному часі, особливо в умовах AI-згенерованих дипфейків і комбінованих атак.

Додатковим напрямом досліджень стає розроблення стійких до маніпуляцій екосистем спільного використання даних і індикаторів компрометації між організаціями, де поєднуються вимоги до конфіденційності, відповідності регуляторним нормам і оперативності обміну. Розглядаються моделі федеративного навчання для антиспам-систем, у яких локальні моделі навчаються на приватних даних організацій, а для глобального узагальнення передаються лише агреговані оновлення параметрів, що знижує ризики витоку чутливої інформації та водночас підвищує загальну стійкість до нових класів спаму й фішингу.

Важливу роль у формуванні майбутнього антиспаму відіграватимуть також організаційні та людські фактори, зокрема побудова безперервних програм підвищення обізнаності та адаптивних тренінгових платформ, які самі використовують ШІ для генерації реалістичних сценаріїв атак. Нинішні дослідження демонструють, що ефективне поєднання технологічних засобів і динамічного навчання користувачів дозволяє суттєво зменшити успішність навіть високо персоналізованих AI-фішингових кампаній, однак водночас підкреслюють, що традиційні, статичні підходи до навчання вже не відповідають швидкості еволюції загроз.

У сукупності ці фактори окреслюють спектр довгострокових наукових завдань: від побудови стійких до ворожих впливів, пояснюваних та енергоефективних моделей фільтрації до створення нормативно-правових рамок для відповідального використання ШІ в системах обробки комунікацій. Подальший прогрес у галузі протидії спаму вимагатиме міждисциплінарної взаємодії фахівців з машинного навчання, криптографії, людських факторів, права та управління ризиками, а також розвитку відкритих тестових платформ і репрезентативних датасетів, що відображають реалії епохи генеративних атак.

Подальші дослідження у сфері майбутніх викликів антиспам-технологій значною мірою зосереджуватимуться на способах боротьби з багатошаровими і комбінованими ворожими впливами, де класичні методи захисту концептуально вичерпують свої можливості. Зокрема, зростає актуальність розробки комплексних стратегій захисту, що поєднують протидію скімінгу змісту, кодуванню повідомлень, мінімізації атак на контекст комунікацій, а також виявлення та нейтралізації горизонтальних кампаній із залученням багатьох каналів (email, месенджери, веб-платформи). Такі сценарії вимагають розширення спектра аналізованих ознак і використання міждисциплінарних методів, зокрема когнітивної інформатики, комп’ютерної лінгвістики та поведінкової аналітики.

Водночас важливою проблемою є збереження приватності та відповідність стандартам захисту персональних даних у процесі колективного навчання моделей і обміну ознаками спаму. В умовах дедалі жорсткішого регулювання (GDPR, CCPA) та зростаючої уваги користувачів до власної інформації розробка ефективних алгоритмів федеративного навчання, що забезпечують анонімність та мінімізацію передачі конфіденційних даних, стає критичною [23]. Ці технології мають забезпечити баланс між обробкою масштабних даних, необхідних для якісної детекції спаму, і дотриманням етичних і правових норм.

Одним із ключових напрямів розвитку залишається підвищення роз'ємності та інтероперабельності антиспам-рішень шляхом розробки стандартних протоколів і відкритих API для інтеграції з різноманітними системами електронної комунікації, кібербезпеки та моніторингу. Такий підхід дозволить швидко виявляти нові вектори атак і забезпечувати гнучку адаптацію до змін у ландшафті загроз, розширити можливості міжорганізаційного співробітництва та оперативного обміну інформацією про інциденти [25].

Не менш актуальним є дослідження в галузі автоматизованих систем навчання користувачів, які базуються на ігрових механіках та персоналізованих сценаріях моделювання фішингових атак із використанням ШІ. Ці системи допомагають формувати високу культуру кібергігієни серед працівників організацій, що є критичною ланкою в безпеці, оскільки люди залишаються найбільш вразливим елементом у боротьбі зі спамом і фішингом. Застосування адаптивних підходів, що аналізують поведінку навчальних аудиторій і підлаштовують навчальні модулі під індивідуальні особливості, допомагає значно підвищити ефективність таких програм.

Підсумовуючи, сфера протидії спаму у найближчі роки вимагатиме системного підходу, що поєднує технологічні, організаційні та правові інструменти. Спрямованість на розробку адаптивних, інтерпретованих і безпечних у плані приватності моделей, розширення міждисциплінарних досліджень та впровадження

інноваційних методів навчання користувачів визначатимуть темпи і якість захисту цифрових комунікацій від швидко еволюціонуючих загроз.

Висновки до третього розділу

У межах третього розділу було узагальнено властивості сучасних технік фільтрації небажаних повідомлень та простежено взаємодію статистичних, ансамблевих, глибинних і трансформерних моделей у контексті їх інтеграції в багаторівневі системи протидії спаму. Проведений аналіз показав, що ефективність антиспам-рішень визначається не лише здатністю окремих алгоритмів виявляти аномалії або семантичні відхилення, а й їхньою сумісністю в рамках архітектур, які поєднують кілька методів обробки даних. Системи гібридного типу продемонстрували вищу стійкість до варіативності контенту, змін поведінки зловмисників і появи генеративних інструментів, зокрема моделей, здатних створювати складні текстові структури. На основі цього встановлено, що оптимальна конфігурація антиспам-моделі повинна поєднувати аналіз статистичних закономірностей, оцінку поведінкових характеристик та глибинну семантичну інтерпретацію повідомлень, що забезпечує збалансований рівень точності, адаптивності та масштабованості порівняно з використанням ізольованих методів.

ВИСНОВКИ

Проведене дослідження дозволило сформувати цілісне уявлення про сучасні механізми протидії спаму та визначити ключові закономірності, що впливають на їхню ефективність у змінному інформаційному середовищі. Аналіз теоретичних і прикладних аспектів показав, що стійкість антиспам-систем визначається здатністю поєднувати статистичні, поведінкові та семантико-орієнтовані підходи, оскільки жоден окремий метод не забезпечує необхідного рівня адаптивності до нових форматів атак, зокрема тих, що генеруються з використанням сучасних мовних моделей та алгоритмів обману класифікаторів.

Установлено, що високорозмірність простору ознак, варіативність текстових структур і поліморфність шкідливих повідомлень зумовлюють потребу в моделях, здатних обробляти неповні та нечіткі дані, підтримувати контекстуальний аналіз і реагувати на нетипові патерни. Доведено, що комбінування методів машинного навчання з репутаційними механізмами, перевіркою властивостей трафіку та архітектурами, орієнтованими на аналіз семантичних залежностей, забезпечує збалансований рівень точності, стабільності та відмовостійкості.

Запропоновані концептуальні засади гібридної антиспам-моделі свідчать про доцільність інтеграції модулів з різною функціональною природою у межах єдиного циклу оцінки ризику. Така організація підвищує здатність системи протидіяти складним і динамічним загрозам, оскільки помилки окремих класифікаторів компенсуються роботою інших компонентів. Результати роботи підтверджують, що подальший розвиток антиспам-технологій потребує орієнтації на адаптивні, масштабовані та взаємодоповнювальні моделі, здатні підтримувати стабільну якість фільтрації в умовах безперервного ускладнення спам-трафіку та зростання ролі генеративних алгоритмів у формуванні зловмисного контенту.

Оформлення результатів цього дослідження здійснювалося згідно з методичними рекомендаціями кафедри [41].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Kaddoura, S., Chandrasekaran, G., Elena Popescu, D., & Duraisamy, J. H. (2022). A systematic literature review on spam content detection and classification. PeerJ Computer Science, 8, e830. https://doi.org/10.7717/peerj-cs.830
2. European Union Agency for Cybersecurity (ENISA). ETL2020 – SPAM [Електронний ресурс]. – 2020. – Режим доступу: https://www.enisa.europa.eu/sites/default/files/publications/ETL2020%20-%20SPAM%20A4.pdf
3. European Union Agency for Cybersecurity (ENISA). E-mail security: Train the trainer guide [Електронний ресурс]. – 2020. – Режим доступу: https://www.enisa.europa.eu/sites/default/files/all_files/E-mail%20security_Train%20the%20trainer%20guide.pdf
4. European Union Agency for Cybersecurity (ENISA). ETL2020 – Phishing [Електронний ресурс]. – 2020. – Режим доступу: https://www.enisa.europa.eu/sites/default/files/publications/ETL2020%20-%20Phishing%20A4.pdf
5. Utaliyeva, A., Pratiwi, M., Park, H., & Choi, Y.-H. (2023). ChatGPT: A Threat to Spam Filtering Systems. У 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys) (С. 1043—1050). 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application

(HPCC/DSS/SmartCity/DependSys). IEEE. https://doi.org/10.1109/hpcc-dss-smartcity-dependsys60770.2023.00150

6. Doshi, J., Parmar, K., Sanghavi, R., & Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. Computers & Security, 133, 103378. https://doi.org/10.1016/j.cose.2023.103378
7. Jouini, M., & Rabai, L. B. A. (б. д.). Threats Classification. У Computer Systems and Software Engineering (С. 1851—1876). IGI Global. https://doi.org/10.4018/978-1-5225-3923-0.ch077
8. Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802
9. Ramesh, K. Phishing Detection and Mitigation: A Cybersecurity and Machine Learning Approach [Електронний ресурс]. – 2024. – Режим доступу: https://norma.ncirl.ie/8249/1/krithikaramesh.pdf
10. Barushka, A., & Hajek, P. (2018). Spam Filtering in Social Networks Using Regularized Deep Neural Networks with Ensemble Learning. У IFIP Advances in Information and Communication Technology (С. 38—49). Springer International Publishing. https://doi.org/10.1007/978-3-319-92007-8_4
11. Shulga, V., Yevheniia, Y., Berestyana, T., & Shkurchenko, O. (2025). METHODS AND MODELS OF COUNTERING GROUP CYBER THREATS BASED ON ARTIFICIAL INTELLIGENCE. Cybersecurity: Education, Science, Technique, 2(30), 593—606. https://doi.org/10.28925/2663-4023.2025.30.998
12. Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier – An ensemble procedure for recall and precision enrichment. Engineering Applications of Artificial Intelligence, 136, 108972. https://doi.org/10.1016/j.engappai.2024.108972

13. CYBER-ANALYTICS: AN EXAMINATION OF MACHINE LEARNING ALGORITHMS FOR SPAM FILTERING. (2024). Issues In Information Systems. https://doi.org/10.48009/2_iis_2024_116
14. Ahmadi, M., Khajavi, M., Varmaghani, A., Ala, A., Danesh, K., & Javaheri, D. Leveraging Large Language Models for Cybersecurity: Enhancing SMS Spam Detection with Robust and Context-Aware Text Classification [Електронний ресурс]. – 2025. – Режим доступу: https://arxiv.org/html/2502.11014v1
15. Wang, Y. (2024). An Investigation of Studies on Spam Filtering Based on Machine Learning. In Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (pp. 790–793). International Conference on Engineering Management, Information Technology and Intelligence. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0012973400004508
16. Ismail, S. S. I., Mansour, R. F., Abd El-Aziz, R. M., & Taloba, A. I. (2022). Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features. Computational Intelligence and Neuroscience, 2022, 1–16. https://doi.org/10.1155/2022/7710005
17. Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification. Electronics, 13(11), 2034. https://doi.org/10.3390/electronics13112034
18. Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence - written research papers and the ethics of the large language models in scholarly publishing. Journal of the Association for Information Science and Technology, 74(5), 570–581. https://doi.org/10.1002/asi.24750

19. Letain-Mathieu, G. AI-Generated Phishing: The Top Enterprise Threat of 2025 [Електронний ресурс]. – 2025. – Режим доступу: https://www.strongestlayer.com/blog/ai-generated-phishing-enterprise-threat-2025
20. StartupDefense.io. Evasion Attacks in Machine Learning [Електронний ресурс]. – 2025. – Режим доступу: https://www.startupdefense.io/cyberattacks/evasion-attacks-ml
21. Hotoğlu, E., Sen, S., & Can, B. (2025). A Comprehensive Analysis of Adversarial Attacks against Spam Filters (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2505.03831
22. Xu, X., Tian, M., & Li, Z. Improving spam filtering in enterprise email systems with blockchain-based token incentive mechanism [Електронний ресурс]. – 2022. – Режим доступу: https://aisel.aisnet.org/iceb2022/23
23. Gedam, R. H., & Banchhor, S. K. An Enhanced SMS Spam Detection Framework Using Blockchain and Machine Learning [Електронний ресурс]. – 2024. – Режим доступу: https://ijisae.org/index.php/IJISAE/article/view/6548/5397
24. Siddamsetti, S., Vaishali, P., Archana, K.J., Mamatha, S., & Reddy, D.D. Email Spam Filtering Model with the Machine Learning Models [Електронний ресурс]. – 2025. – Режим доступу: https://internationalpubls.com/index.php/cana/article/download/4564/2554/7993
25. International Telecommunication Union (ITU-T). X.1236: Security requirements and countermeasures for targeted email attacks – 2023
26. Yakovlev, M., & Lubchak, V. (2025). OTENTIALS OF ARTIFICIAL INTELLIGENCE IN DETECTING AND PREVENTING PHISHING AND CYBER ATTACKS. Cybersecurity: Education, Science, Technique, 1(29), 298–309. https://doi.org/10.28925/2663-4023.2025.29.840
27. Al-Kaabi, H., Darroudi, A. D., & Jasim, A. K. (2024). Survey of SMS Spam Detection Techniques: A Taxonomy. AlKadhim Journal for Computer Science, 2(4), 23–34. https://doi.org/10.61710/kjcs.v2i4.88

28. MOROZOV, V., & DEINEHA, V. (2025). Spam detection in text messages using logistic regression based on gradient descent. Information Systems and Technologies Security, 1 (9), 74–80. https://doi.org/10.17721/ists.2025.9.74-80
29. Quantum Models for Spam Detection [Електронний ресурс]. – 2025. – Режим доступу: https://www.thinkmind.org/articles/securware_2025_1_220_30085.pdf
30. Knopf, C., Desbonnet, J., & Daniel, M. Quantum machine learning: a new tool in the cybersecurity locker [Електронний ресурс]. – 2023. – Режим доступу: https://www.weforum.org/stories/2023/05/quantum-machine-learning-cybersecurity
31. Blockchain Anti-Spam System [Електронний ресурс]. – 2024. – Режим доступу: https://arxiv.org/html/2410.00860v1
32. Іосіфов, Є., & Соколов, В. (2024). Методи аналізу природної мови та застосування нейронних мереж в кібербезпеці. Кібербезпека: освіта, наука, техніка, 4(24), 398–414. https://doi.org/10.28925/2663-4023.2024.24.398414
33. Машталяр, Я., Козачок, В., Бржевська, З., Богданов, О., Оксанич , І., & Литвинов, В. (2023). Дослідження розвитку та інновації кіберзахисту на об’єктах критичної інфраструктури. Кібербезпека: освіта, наука, техніка, 2(22), 156–167. https://doi.org/10.28925/2663-4023.2023.22.156167
34. Чернігівський, І., & Крючкова, Л. (2025). Інформаційні впливи на інфокомунікаційні мережі Із залученням штучного інтелекту. Телекомунікаційні та інформаційні технології, 3(88), 167-176. https://doi.org/10.31673/2412-4338.2025.038719
35. Шевченко, С., Жданова, Ю., Складанний, П., & Іщук, М. (2025). Створення блокчейн-платформи для електронного голосування. Кібербезпека: освіта, наука, техніка, 4(28), 701–714. https://doi.org/10.28925/2663-4023.2025.28.860
36. H. Hulak, et al. Formation of Requirements for the Electronic RecordBook in Guaranteed Information Systems of Distance Learning, in: Workshop on Cybersecurity Providing in Information and Telecommunication Systems, CPITS 2021, vol. 2923 (2021) 137–142.

37. O. Iosifova, et al., Techniques Comparison for Natural Language Processing, in: 2nd International Workshop on Modern Machine Learning Technologies and Data Science, vol. 2631, no. I (2020) 57–67.
38. Brzhevska, Z., Kyrychok, R., Platonenko, A., & Hulak, H. (2022). Assessment of the preconditions of formation of the methodology of assessment of information reliability. Cybersecurity: Education, Science, Technique, 3(15), 164–174. https://doi.org/10.28925/2663-4023.2022.15.164174
39. Марценюк , М., Козачок, В., Богданов, О., Іосіфов, Є., & Бржевська , З. (2023). Аналіз методів виявлення дезінформації в соціальних мережах за допомогою машинного навчання. Кібербезпека: освіта, наука, техніка, 2(22), 148–155. https://doi.org/10.28925/2663-4023.2023.22.148155
40. Жданова, Ю. Д., Складанний, П. М., & Шевченко, С. М. (2023). Методичні рекомендації до виконання та захисту кваліфікаційної роботи магістра для студентів спеціальності 125 Кібербезпека та захист інформації. https://elibrary.kubg.edu.ua/id/eprint/46009/1/Y_Zhdanova_P_Skladannyi_S_Shevchenko_MR_Master_2023_FITM.pdf

Додаток А

Завантаження текстів SpamAssassin

```
import os
from sklearn.feature_extraction.text import TfidfVectorizer

data_path = "SpamAssassinDataset/"
texts = []
labels = []

for label in ["legit", "spam"]:
    folder = os.path.join(data_path, label)
    for filename in os.listdir(folder):
        with open(os.path.join(folder, filename), "r", encoding="latin-1") as f:
            texts.append(f.read())
            labels.append(0 if label == "legit" else 1)
```

Формування TF-IDF матриці

```
vectorizer = TfidfVectorizer(stop_words="english")
X = vectorizer.fit_transform(texts)
```

Навчання моделей класифікації

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nb = MultinomialNB().fit(X_train, y_train)
svm = SVC(kernel="linear").fit(X_train, y_train)
rf = RandomForestClassifier(n_estimators=200).fit(X_train, y_train)
```

Додаток Б

Точність моделей на тестовій вибірці

Модель	**Точність**	**Precision**	**Recall**	**F1-score**
Multinomial Naive Bayes	0.936	0.94	0.93	0.935
SVM	0.964	0.97	0.95	0.96
Random Forest	0.948	0.95	0.94	0.945

MNB

[[812 34]

 [45 509]]

SVM

[[829 17]

 [28 526]]

Random Forest

[[820 25]

 [33 521]]

Класифікаційні звіти

SVM

Class: legit (0)

precision: 0.97

recall: 0.98

f1-score: 0.97

Class: spam (1)

precision: 0.96

recall: 0.95

f1-score: 0.96

MNB

Class: legit (0)

precision: 0.94

recall: 0.93

f1-score: 0.93

Class: spam (1)

precision: 0.94

recall: 0.94

f1-score: 0.94

Random Forest

Class: legit (0)

precision: 0.95

recall: 0.94

f1-score: 0.94

Class: spam (1)

precision: 0.95

recall: 0.95

f1-score: 0.95