Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод класифікації конфіденційної інформації із застосуванням машинного навчання*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент групи КН-21-2* — Підпис — *Богдан ПАЛІЙЧУК*
Група виконавця — Ім'я, ПРІЗВИЩЕ

Керівник: *ст. викладач каф. КН* — Підпис — *Тетяна СКРИПНИК*
Науковий ступінь, посада — Ім'я, ПРІЗВИЩЕ

Нормоконтроль: *к.т.н. доц. каф. КН* — Підпис — *Руслан БАГРІЙ*
Науковий ступінь, посада — Ім'я, ПРІЗВИЩЕ

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис — *Олександр БАРМАК*
Ім'я, ПРІЗВИЩЕ

18 *06* 2025 р.

Хмельницький 2025

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*10*» *02* *2025* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод класифікації конфіденційної інформації із застосуванням машинного навчання»*

2. Завдання видано студенту *Богдану ПАЛІЙЧУКУ*
(Ім'я, прізвище)

3. Керівник роботи *ст. викладач кафедри КН Тетяна СКРИПНИК*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*07*» *02* *2025* р. № *23*

5. Дата видачі завдання студентці: «*10*» *02* *2025* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Метою кваліфікаційної роботи є підвищення точності класифікації конфіденційної інформації із застосуванням машинного навчання. Перелік задач: провести аналіз предметної області задач класифікації текстової інформації; дослідити існуючі методи для класифікації конфіденційної інформації в текстах за допомогою машинного навчання; розробити метод для ідентифікації конфіденційної інформації в текстових даних; провести експериментальні дослідження точності розробленого методу для класифікації конфіденційної інформації.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи бакалавра з керівником*	*січень 2025*	*виконано*
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*лютий 2025*	*виконано*
3	*Робота над розділом 1 – Характеристика предметної області та постановка задачі*	*березень 2025*	*виконано*
4	*Робота над розділом 2 – Метод класифікації конфіденційної інформації із застосуванням машинного навчання*	*квітень 2025*	*виконано*
5	*Робота над розділом 3 – Експериментальна перевірка методу класифікації конфіденційної інформації із застосуванням машинного навчання*	*травень 2025*	*виконано*
6	*Оформлення пояснювальної записки згідно вимог*	*травень 2025*	*виконано*
7	*Попередній захист кваліфікаційної роботи бакалавра*	*травень 2025*	*виконано*
8	*Захист кваліфікаційної роботи бакалавра на засіданні Екзаменаційної комісії*	*червень 2025*	*виконано*

Виконавець: *тудент групи КН-21-2* (Група виконавця) _____ (Підпис) *Богдан ПАЛІЙЧУК* (Ім'я, ПРІЗВИЩЕ)

Керівник: *ст. викладач каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Тетяна СКРИПНИК* (Ім'я, ПРІЗВИЩЕ)

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод класифікації конфіденційної інформації із застосуванням машинного навчання».*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-21-2 Богдан ПАЛІЙЧУК.*

Керівник кваліфікаційної роботи бакалавра: *ст. викладач каф. КН Тетяна СКРИПНИК.*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
52	14	3	41	3

Метою кваліфікаційної роботи є підвищення точності класифікації конфіденційної інформації із застосуванням машинного навчання.

Запропонований підхід дає змогу автоматизовано виявляти та класифікувати конфіденційні відомості в текстах, визначати їх ключові характеристики, а також здійснювати моніторинг точності заходів із захисту таких даних у процесі їх використання.

Результати виконаної кваліфікаційної роботи бакалавра свідчать про високу точність розробленого методу класифікації конфіденційної інформації із застосуванням машинного навчання.

Ключові слова: машинне навчання, конфіденційна інформація, класифікація, інформаційна безпека, згладжування Лапласа, наївний баєсівський класифікатор, метод опорних векторів, F1-міра.

Виконавець: *студент групи КН-21-2* _____ *Богдан ПАЛІЙЧУК*

Група виконавця | Підпис | Ім'я, ПРІЗВИЩЕ

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
ML	ML Machine Learning (Машинне навчання)
НМ	НМ Нейронна мережа
КІ	КІ Конфіденційна інформація
БД	БД База даних
НТ	НТ Непублічна тема
ТКІ	ТКІ Техніки класифікації інформації

Вступ

Кваліфікаційна робота бакалавра присвячена розробці методу класифікації конфіденційної інформації із застосуванням машинного навчання. Запропонований підхід дає змогу автоматизовано виявляти та класифікувати конфіденційні відомості в текстах, визначати їх ключові характеристики, а також здійснювати моніторинг точності заходів із захисту таких даних у процесі їх використання. Це сприятиме забезпеченню високого рівня безпеки критично важливої інформації – фінансових, медичних, персональних даних і об'єктів інтелектуальної власності – від несанкціонованого доступу та витоку.

Актуальність. У сучасному цифровому середовищі обсяги інформації стрімко зростають, і значна її частина містить конфіденційні дані. Забезпечення захисту таких даних є одним із пріоритетних завдань як для організацій, так і для окремих користувачів. До конфіденційної інформації належать, зокрема, фінансові звіти, медична документація, персональні дані, об'єкти інтелектуальної власності – усі ці відомості потребують захисту від несанкціонованого доступу та витоку.

З огляду на зростаючу складність та обсяги даних, традиційні підходи до їх класифікації та аналізу вже не забезпечують належної точності. У цьому контексті використання методів машинного навчання стає актуальним і перспективним напрямом. Такі методи дають змогу створювати автоматизовані системи для розпізнавання та класифікації конфіденційної інформації в текстах, навіть за відсутності явних маркерів.

Моделі машинного навчання, зокрема нейронні мережі, мають здатність виявляти приховані закономірності та контексти, що дозволяє точніше визначати й захищати конфіденційні дані. Їх впровадження сприяє зменшенню впливу людського фактора, підвищенню точності обробки інформації, а також прискоренню процесу аналізу.

Крім того, дотримання міжнародних норм і стандартів, таких як ЗРЗД у Європі чи HIPAA у США, вимагає від компаній впровадження точних

інструментів захисту інформації. У зв’язку з цим методи машинного навчання виступають дієвим засобом забезпечення інформаційної безпеки.

Зростання кількості цілеспрямованих кібератак ще більше підсилює потребу в розробці сучасних рішень для виявлення та захисту конфіденційну інформації. Таким чином, тема кваліфікаційної роботи є актуальною, оскільки вона спрямована на підвищення рівня безпеки обробки текстових даних за допомогою інтелектуальних технологій.

Об’єктом дослідження є процес класифікації конфіденційної інформації.

Предметом дослідження є методи машинного навчання для аналізу текстових даних.

Метою кваліфікаційної роботи бакалавра є підвищення точності класифікації конфіденційної інформації із застосуванням машинного навчання.

Для досягнення поставленої мети необхідно реалізувати такі задачі:

– провести аналіз предметної області задач класифікації текстової інформації;

– дослідити існуючі методи для класифікації конфіденційної інформації в текстах за допомогою машинного навчання;

– розробити метод для ідентифікації конфіденційної інформації в текстових даних;

– провести експериментальні дослідження точності розробленого методу для класифікації конфіденційної інформації.

Розділ 1 Характеристика предметної області, аналіз методів та реалізацій

1.1 Конфіденційна інформація та необхідність її визначення і класифікації

У зв’язку зі зростанням обсягів неструктурованих даних, методи класифікації залишаються актуальним напрямом досліджень. Останні досягнення в машинному та глибокому навчанні, зокрема нейронних мережах, значно розширили можливості аналізу даних. Сучасні класифікаційні методи дають змогу точно обробляти великі масиви інформації та прогнозувати поведінкові моделі, що особливо корисно для бізнесу при аналізі клієнтських уподобань.

Однак, як показують останні дослідження, все більше людей усвідомлюють масштаби збирання їхніх особистих даних, але мало хто знає, як саме ці дані використовуються.

Загальний регламент захисту даних (ЗРЗД) ЄС зобов’язує компанії уважно ставитися до обробки персональних даних, що дозволяють прямо чи опосередковано ідентифікувати особу. Однак дослідження показують, що навіть анонімні дані можуть бути використані для ідентифікації, якщо їх поєднати з іншими джерелами. У світі масового збору інформації це створює додаткові виклики щодо дотримання вимог регламенту та забезпечення конфіденційності.

У світлі цього стає очевидною необхідність подальших досліджень та розробки точних методів класифікації конфіденційної інформації. Такі методи мають бути не лише надійними та точними, але й відповідати вимогам законодавства щодо захисту даних.

Зокрема, перспективним є використання алгоритмів машинного навчання та NLP для автоматизованого виявлення конфіденційної інформації в текстах. Також важливими напрямами є розробка методів шифрування, анонімізації та аналізу впливу класифікаційних систем на дотримання стандартів захисту даних, таких як ЗРЗД.

Після резонансних подій, що загрожували приватності особистих даних, стало очевидно, що існуючі підходи до класифікації конфіденційної інформації потребують вдосконалення. Зростає потреба у нових методах, здатних своєчасно виявляти та захищати конфіденційні дані від несанкціонованого доступу.

Розвиток точних систем класифікації є критично важливим для гарантування безпеки та приватності в умовах цифрової обробки великих обсягів інформації.

1.2 Аналіз наукових досліджень для класифікації конфіденційних даних

У цифрову епоху, коли обсяги даних стрімко зростають, захист конфіденційної інформації стає надзвичайно важливим. Точні методи класифікації допомагають виявляти та захищати такі дані від несанкціонованого доступу. Машинне навчання є ключовим інструментом для автоматизованого аналізу великих обсягів інформації та виявлення складних закономірностей. Наукові дослідження з класифікації конфіденційних даних охоплюють комп’ютерні науки, інформаційну безпеку та обробку природної мови.

Дослідження пропонує метод захисту конфіденційності навчальних даних у машинному навчанні за допомогою диференційної приватності [1]. Цей підхід гарантує, що особисті дані не будуть розкриті або використані для ідентифікації осіб. Метод полягає у додаванні шуму або інших змін до даних чи параметрів моделі під час тренування, що знижує ризик витоку інформації без втрати точності моделі.

Стаття пропонує огляд технік збереження конфіденційності в інтелектуальному аналізі даних, включаючи анонімізацію, спотворення, криптографію, нечітку логіку та нейронні мережі [2]. Автори розділяють методи на централізовані (з довіреним сервером) та розподілені (де учасники приховують свої дані), порівнюють їх точність для завдань класифікації, кластеризації та виявлення аномалій. Вказано на компроміс між захистом

приватності та втратою інформації. Стаття підкреслює важливість захисту даних у сферах охорони здоров'я, фармацевтики та безпеки, а також окреслює перспективи подальших досліджень.

У роботі [3] наведено огляд методів захисту конфіденційності на різних етапах життєвого циклу даних: збір, публікація, розподіл та аналіз. Для кожного етапу описано основні техніки, їх переваги й недоліки, зокрема рандомізацію, анонімізацію (k-анонімність, l-різноманітність, t-близькість, диференційну приватність), приховування правил, контроль запитів та розподілені протоколи. Розглянуто метрики для оцінки конфіденційності й корисності даних, а також приклади застосування в медицині, банківській сфері, кібербезпеці, торгівлі та транспорті. Огляд є структурованим і вичерпним, проте потребує глибшого критичного аналізу методів і перспектив їх розвитку. Вибір оптимального методу залежить від типу і обсягу даних та вимог до точності.

Дослідження зосереджені на застосуванні машинного навчання для класифікації конфіденційних даних та захисту приватності під час навчання моделей [4–9]. Описано різні методи, зокрема диференційну приватність, анонімізацію, рандомізацію й криптографію [10–14], що забезпечують безпеку на різних етапах життєвого циклу даних.

Ці роботи показують широкий спектр підходів до класифікації конфіденційної інформації та важливість захисту даних у цьому процесі.

Більшість досліджень фокусуються на аналізі великих обсягів неструктурованих текстів (новини, соцмережі, пошта), що дозволяє автоматизувати обробку та прискорити доступ до важливої інформації. Водночас частина робіт розглядає класифікацію структурованих документів — звітів, медичних записів тощо — актуальну для галузей із формалізованими даними.

Дослідження класифікації текстів охоплюють різні підходи: експертні правила, класичні алгоритми машинного навчання (наївний Баєс, метод опорних векторів, дерева рішень) [15–17] та глибоке навчання з нейронними мережами [21–30]. Також вивчають точність різних функцій втрат і метрик (precision, recall, F-міра, ROC, AUC) [31–34].

Особлива увага приділяється методам аналізу неструктурованих даних — текстових повідомлень, електронної пошти, веб-сторінок і соцмереж.

Окрім теоретичних досліджень, класифікація текстів широко застосовується на практиці: для аналізу настроїв у соцмережах [35–38], обробки електронної пошти, класифікації новин, виявлення спаму та сортування документів [39–40].

Аналіз настроїв у соцмережах — ключове практичне застосування, що дозволяє автоматично визначати емоції користувачів у великих обсягах тексту. Це допомагає бізнесу оцінювати реакцію аудиторії, виявляти проблеми та покращувати маркетингові стратегії.

Компанії використовують аналіз настроїв для виявлення та реагування на скарги, пропозиції й питання клієнтів у соцмережах. Це допомагає відстежувати зміни вподобань аудиторії, підтримувати розробку нових продуктів і швидко реагувати на кризові ситуації. Водночас важливо враховувати етичні питання та захист приватності користувачів.

Автоматична обробка електронної пошти — ще одна важлива сфера застосування класифікації текстів, що дозволяє точно фільтрувати та сортувати вхідні повідомлення за допомогою машинного навчання.

Класифікація текстових повідомлень допомагає відокремлювати спам від корисної пошти, захищаючи вхідні скриньки від небажаних листів. Також здійснюється сортування листів за категоріями, наприклад, важливість чи тема, для полегшення їх обробки.

Автоматичне виявлення важливих повідомлень і створення відповідей на основі їхнього змісту допомагає користувачам швидко реагувати та керувати великою кількістю електронної пошти, підвищуючи продуктивність.

Класифікація новинних статей — важлива сфера застосування, що допомагає автоматично визначати тематику матеріалів. Це полегшує користувачам швидкий пошук та отримання інформації за інтересами. Також класифікація дозволяє фільтрувати новини за темою, датою, джерелом чи

автором, роблячи пошук зручнішим. Вона допомагає швидко виявляти та відстежувати ключові події у різних сферах.

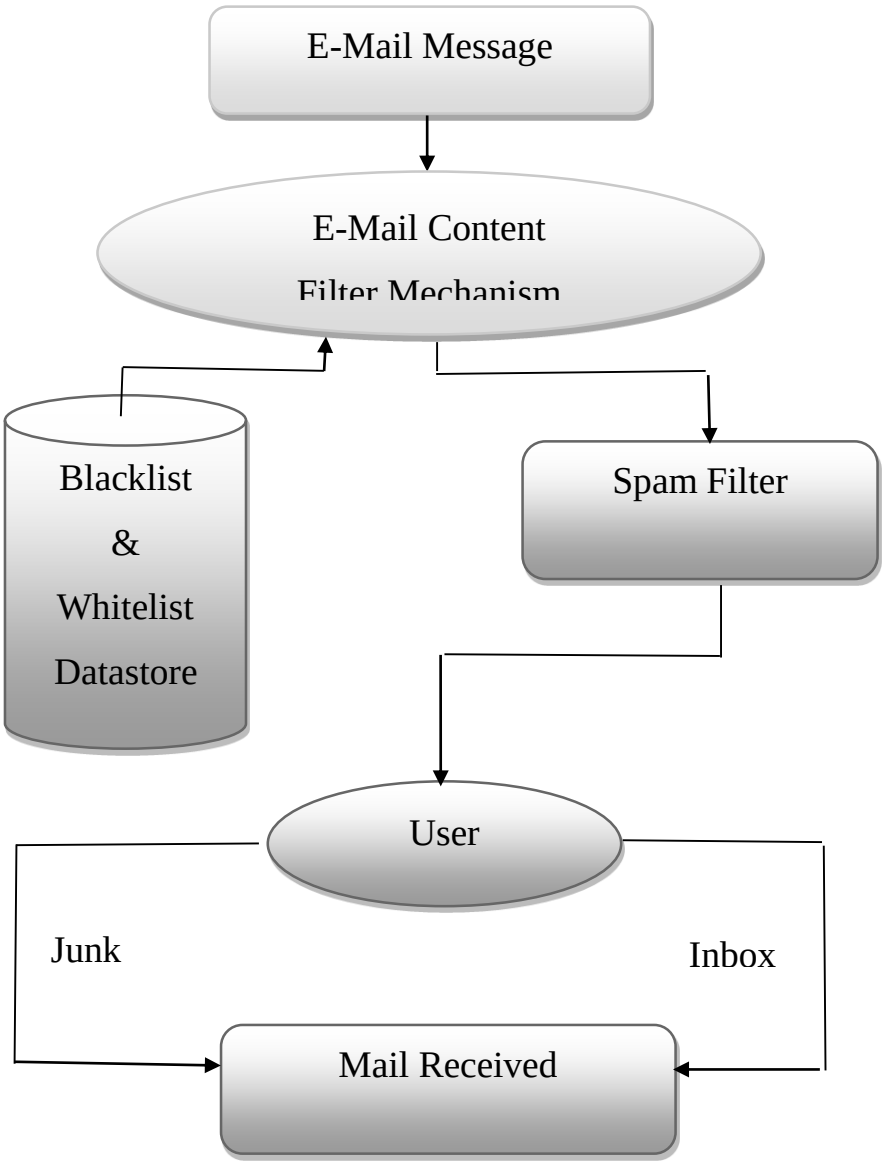


Рисунок 1.1 – Модель процедури спам-фільтра електронної пошти

Класифікація новин за тоном і сентиментом допомагає зрозуміти громадське сприйняття та емоції щодо інформації. Автоматична перевірка фактів спрямована на виявлення і видалення дезінформації та фейкових новин. Ці методи застосовуються на новинних платформах, у соцмережах і пошукових системах для підвищення якості та актуальності контенту, зменшуючи час пошуку потрібної інформації.

Виявлення спаму — ключове завдання в обробці електронної пошти та комунікацій, щоб уникнути потоку небажаних повідомлень. Основні методи базуються на правилах і евристиках, що ідентифікують спам за характерними ознаками тексту, ключовими словами або підозрілими адресами відправників.

Алгоритми машинного навчання, такі як наївний Баєс, метод опорних векторів і нейронні мережі, автоматично визначають спам за характеристиками тексту, відправника та структури повідомлення. Евристики виявляють спам за відомими патернами, наприклад, частотою відправлення або специфічним вмістом. Комбінація методів підвищує точність і зменшує помилки. Виявлення спаму покращує користувацький досвід і безпеку на платформах електронної пошти, соцмережах і форумах.

Автоматичне сортування документів визначає їхню тему чи категорію за вмістом, що допомагає швидко організувати великі обсяги інформації. Також оцінюється важливість документів для пріоритезації подальшої обробки.

Документи автоматично розпізнають і сортують за типом (тексти, таблиці, презентації). Класифікація фільтрує документи за ключовими словами та параметрами, а також оновлює метадані для полегшення пошуку. Вона використовується для персоналізованих рекомендацій на основі поведінки користувачів.

Класифікація допомагає виявляти шаблони та тренди в великих даних, аналізувати соцмережі для моніторингу брендів у реальному часі, автоматизувати обробку юридичних документів і медіа-моніторинг.

Крім того, класифікація застосовується у медицині, фінансах, аналізі веб-контенту та автоматичному тегуванні, що робить обробку текстів швидшою і точнішою.

Дослідження в цій галузі спрямовані на розвиток методів для глибшого і ширшого аналізу текстової інформації у різних сферах.

1.3 Методи машинного навчання для класифікації конфіденційних даних

Одним із ключових викликів у створенні моделі машинного навчання є забезпечення якісних та релевантних даних. Зазвичай використовуються два типи наборів: навчальні — для того, щоб модель вивчила закономірності, та тестові — для перевірки її точності на нових даних.

Дані повинні відповідати задачі, яку розв'язує модель. У процесі навчання модель підлаштовує свої параметри, щоб максимально точно прогнозувати результати на основі вхідної інформації.

Після навчання модель перевіряють на тестових даних, які не використовувалися раніше. Це дозволяє оцінити, як добре вона прогнозує результати на новій інформації, порівнюючи її відповіді з правильними.

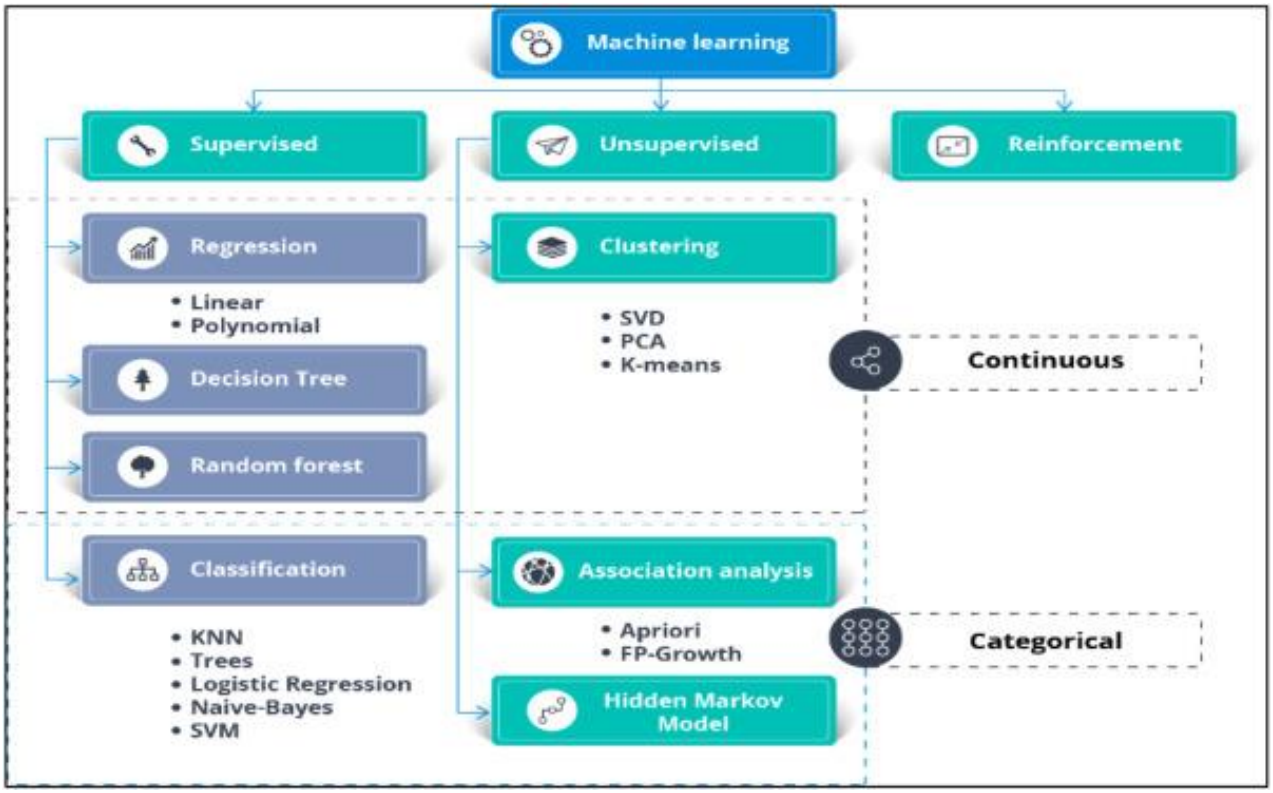


Рисунок 1.2 – Методи машинного навчання [41]

Процес навчання моделі повторюють кілька разів, змінюючи параметри або алгоритми, щоб підвищити її точність. У результаті формується модель, здатна робити прогнози на нових даних.

Головна мета — узагальнення знань і стабільна робота на подібних, але нових задачах. Проте можуть виникати проблеми:

– *Перенавчання* — модель добре працює на навчальних даних, але погано на нових, бо запам’ятовує дані замість узагальнення.

– *Недонавчання* — модель не засвоює навіть навчальні дані, зазвичай через надто просту архітектуру або малий обсяг даних.

Метод опорних векторів (SVM) — це алгоритм контрольованого навчання, що використовується для класифікації текстів. Його мета — знайти оптимальну межу, яка розділяє простір ознак між різними класами. Цей метод точно визначає границі між категоріями, що робить його корисним у текстовій класифікації.

Комп’ютери добре працюють зі структурованими даними, але людська мова є неструктурованою та багатозначною, що ускладнює її обробку. Щоб інтерпретувати текст, комп’ютеру потрібно розуміти значення слів у контексті. Обробка природної мови (NLP) фокусується саме на семантиці, на відміну від розпізнавання мовлення, яке аналізує звук.

1.4 Мета, задачі роботи

Метою кваліфікаційної роботи є підвищення точності класифікації конфіденційної інформації із застосуванням машинного навчання.

Задачі:

– провести аналіз предметної області задач класифікації текстової інформації;

– дослідити існуючі методи для класифікації конфіденційної інформації в текстах за допомогою машинного навчання;

– розробити метод для ідентифікації конфіденційної інформації в текстових даних;

– провести експериментальні дослідження точності розробленого методу для класифікації конфіденційної інформації.

Розділ 2 Метод класифікації конфіденційної інформації із застосуванням машинного навчання

2.1 Структура методу класифікації конфіденційної інформації

Метод класифікації конфіденційної інформації – це систематичний процес виявлення, аналізу та присвоєння рівнів конфіденційності до інформації з метою забезпечення її належного захисту відповідно до рівня ризику та впливу в разі її розголошення, втрати або модифікації.

На рисунку 2.1 зображено блок-схему архітектури методології методу класифікації конфіденційної інформації складається з п’яти етапів, що утворюють ітераційний, структурований процес.

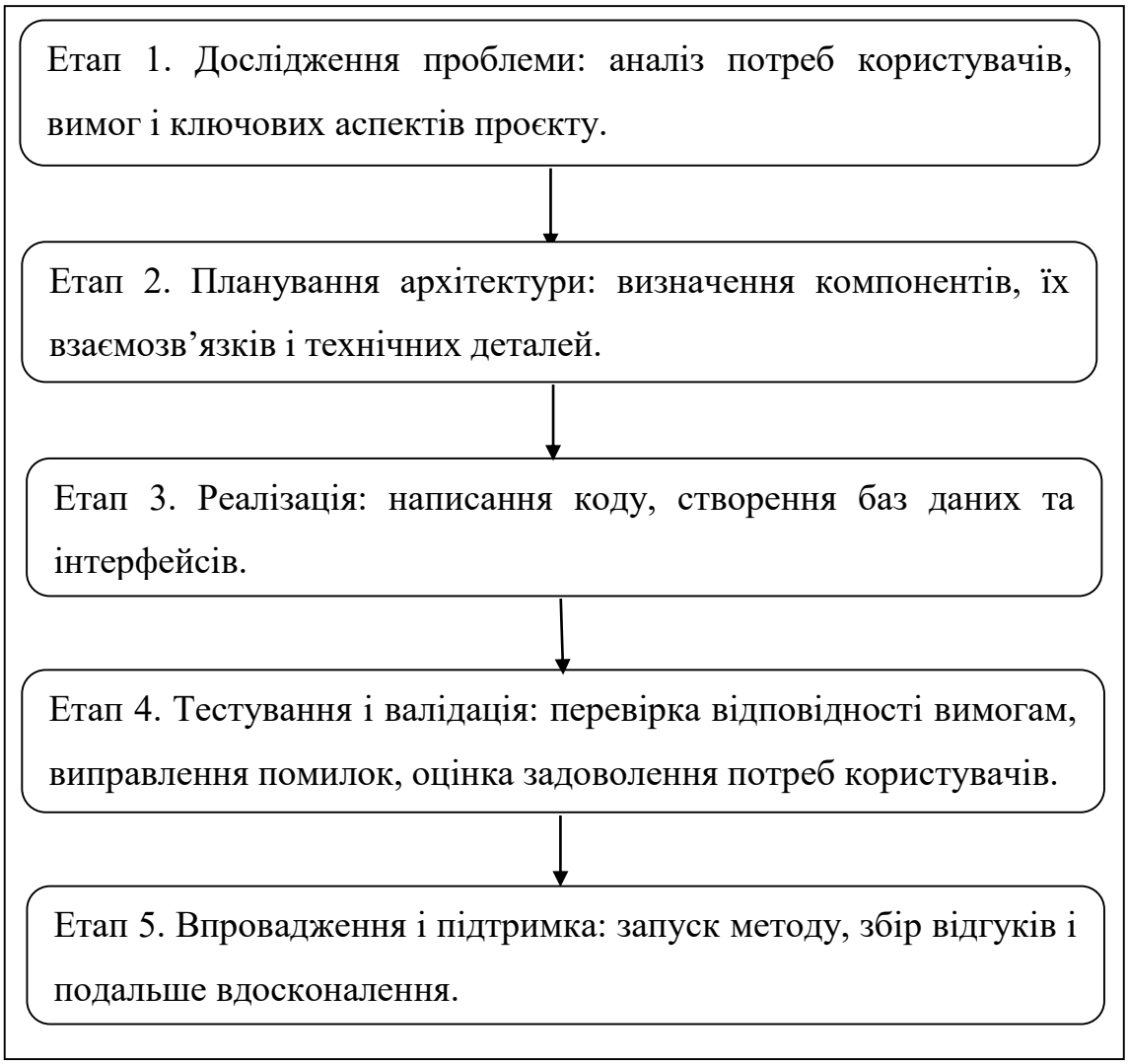


Рисунок 2.1 – Архітектура методу класифікації конфіденційної інформації

Перший етап включає глибокий аналіз проблеми, виявлення викликів і потреб різних користувачів, що формує основу для вимог і подальшої розробки.

На цьому етапі визначається архітектура методу — створюються схеми, компоненти та їх взаємозв’язки з урахуванням потреб користувачів і технічних можливостей. Обирається оптимальний варіант для точної реалізації.

Також розробляються плани впровадження: розподіл завдань, стратегії розробки та необхідні технічні засоби.

Після цього починається безпосередня реалізація: написання коду, створення баз даних і компонентів методу. Розробники створюють функціональні модулі та дотримуються стандартів програмування для якісної роботи системи.

Створення баз даних включає проєктування схем, таблиць, полів і зв’язків, а також розробку запитів і процедур для точного доступу до даних і зручності користування методом.

Після розробки проводиться тестування: модульне, функціональне, інтеграційне та системне, щоб виявити й виправити помилки. Потім виконується валідація, яка перевіряє відповідність методу потребам користувачів у реальних умовах.

На фінальному етапі метод впроваджується — встановлюється, налаштовується, користувачів навчають роботі з ним, а також забезпечується техпідтримка. Після цього проводиться остаточна оцінка роботи системи за стабільністю, безпекою і масштабованістю.

2.2 Метод класифікації конфіденційної інформації із застосуванням машинного навчання

Метою методу є автоматичне виявлення та класифікація конфіденційних даних у великих масивах інформації. Моделі машинного навчання дозволяють з високою точністю розрізняти конфіденційну та неконфіденційну інформацію, що сприяє підвищенню захисту даних і дотриманню вимог конфіденційності.

Такий підхід є ключовим для забезпечення інформаційної безпеки. Реалізація методу передбачає низку типових етапів.

Початковим етапом є збір і підготовка набору даних із прикладами як конфіденційної, так і неконфіденційної інформації (текст, зображення тощо). Спершу слід чітко визначити, які дані вважаються конфіденційними (наприклад, ПІБ, адреса, email) і які — ні. Далі потрібно зібрати репрезентативні приклади обох типів. Якщо використовуються реальні особисті дані, необхідно провести анонімізацію або псевдонімізацію для захисту приватності.

Баланс і якість даних. Під час збору важливо забезпечити приблизно рівну кількість конфіденційних і неконфіденційних прикладів, щоб уникнути перекосу під час навчання моделі. Перед тренуванням слід перевірити якість даних — усунути помилки й переконатися у їх відповідності поставленій задачі.

Вибір моделі. Для класифікації даних обирають відповідну модель машинного навчання (наприклад, логістична регресія, SVM, наївний Байєс, нейромережі). Вибір залежить від типу даних, їх обсягу, доступних ресурсів та вимог до точності.

Наївний Баєс — простий і швидкий, точний для великих обсягів даних, але припускає незалежність ознак, що не завжди коректно.

Логістична регресія — легка в реалізації та інтерпретації, підходить для лінійно роздільних даних, але не справляється зі складними залежностями.

Метод опорних векторів (SVM) — точний, але вимагає ретельного налаштування параметрів.

Глибокі нейронні мережі — потужні для складних задач і великих даних, однак потребують багато ресурсів і обережного налаштування для уникнення перенавчання.

Модель навчається на підготовленому наборі даних, аналізуючи патерни, що дозволяють відрізнити конфіденційну інформацію від неконфіденційної. Під час навчання вона коригує свої параметри (наприклад, за допомогою градієнтного спуску), щоб зменшити помилки класифікації.

Головне — досягти балансу між перенавчанням (коли модель надто точно запам’ятовує навчальні дані) і недонавчанням (коли не виявляє суттєвих залежностей).

Після навчання модель тестують на нових даних, щоб оцінити її точність і готовність до практичного використання.

Оцінка моделі. Після навчання модель перевіряють на незалежному наборі даних за допомогою метрик:

Accuracy — загальна точність класифікації;

Precision — частка правильних серед усіх передбачених позитивних прикладів;

Recall — частка знайдених позитивних прикладів серед усіх справжніх;

F1-оцінка — баланс між precision і recall.

Вибір метрик залежить від задачі: у деяких випадках важливіше мінімізувати пропущені позитивні приклади, в інших — хибнопозитивні. За потреби модель донавчають або коригують її параметри для покращення результатів.

Налаштування параметрів. Якщо модель показує незадовільні результати, слід налаштувати її гіперпараметри — значення, що впливають на роботу, але не навчаються автоматично (наприклад, параметри ядра чи регуляризації в SVM).

Для пошуку оптимальних параметрів використовують перехресну валідацію, що дозволяє уникнути перенавчання. Оптимізацію здійснюють через методи, як-от випадковий пошук або градієнтний спуск.

Після налаштування модель повторно оцінюють на незалежному наборі даних для перевірки покращення.

Валідація та розгортання. Налаштування параметрів — ітеративний процес, що вимагає врахування як технічних, так і прикладних аспектів задачі. Після успішного навчання модель проходить валідацію на нових, реальних даних для перевірки її стабільності та надійності.

Далі відбувається розгортання — інтеграція моделі у виробниче середовище чи систему. Важливо забезпечити моніторинг продуктивності, реагування на зміни у вхідних даних та своєчасне оновлення. Підтримка та вдосконалення моделі є безперервною частиною її життєвого циклу.

Цей процес можна подати у вигляді блок-схеми з послідовними етапами та стрілками, що вказують напрям переходу між ними.

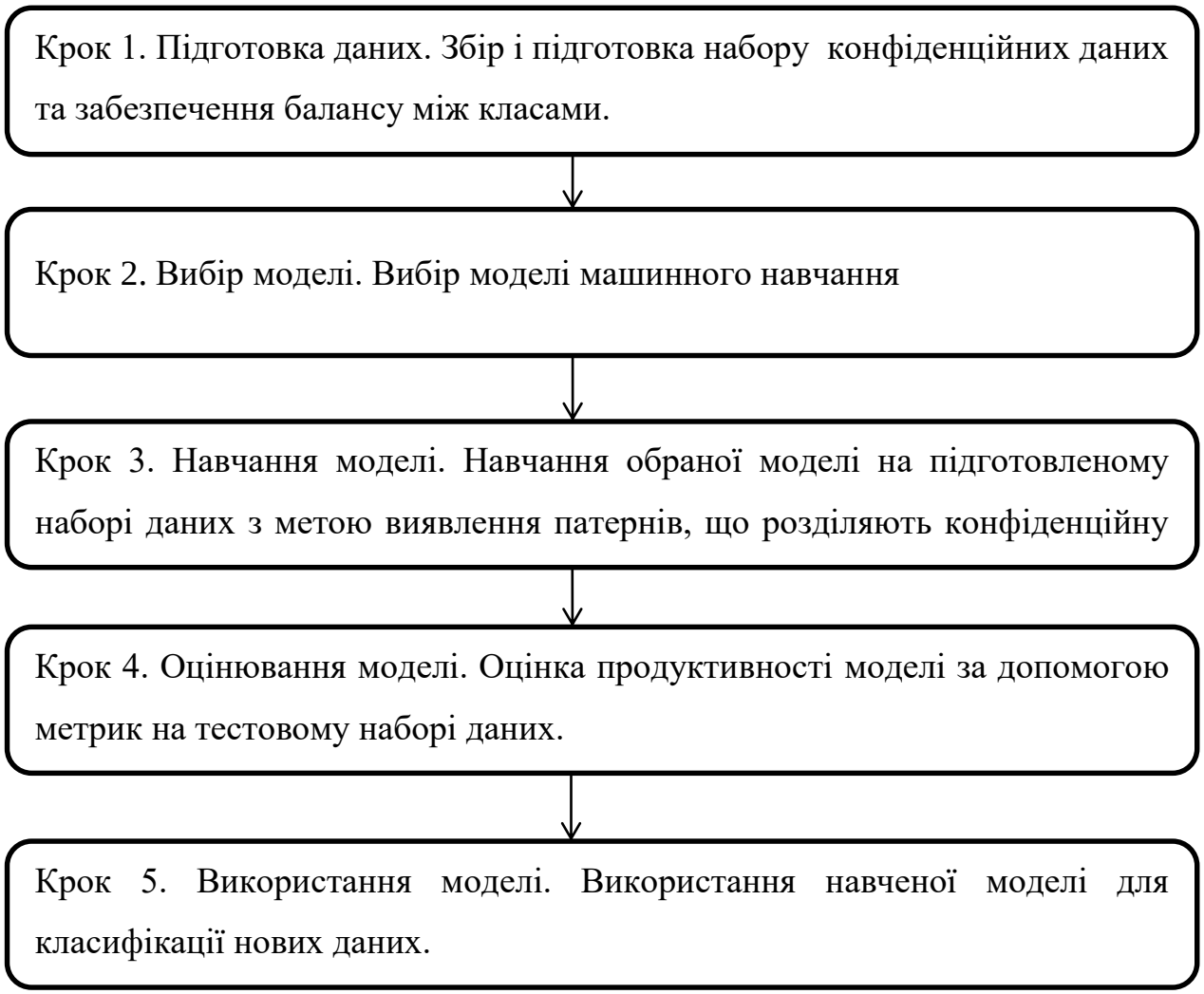


Рисунок 2.2 – Загальний опис метода класифікації конфіденційної інформації із застосуванням машинного навчання

Ця діаграма показує структуру та взаємозв’язки в наборі даних. Центральним елементом є "Набір даних", який поділяється на три основні частини:

Тренувальні дані — використовуються для навчання моделі, виявлення закономірностей і побудови алгоритмів.

Тестові дані — служать для перевірки точності та надійності моделі. Вони не використовуються в процесі навчання.

Кросвалідація — метод оцінки, що застосовується до тренувальних і тестових даних для підвищення достовірності результатів. Найпоширеніший підхід — K-fold cross-validation, коли дані ділять на K частин і по черзі використовують кожну як тестову.

Схема на рисунку 2.3 ілюструє послідовний і надійний підхід до підготовки даних для навчання та оцінки моделей.

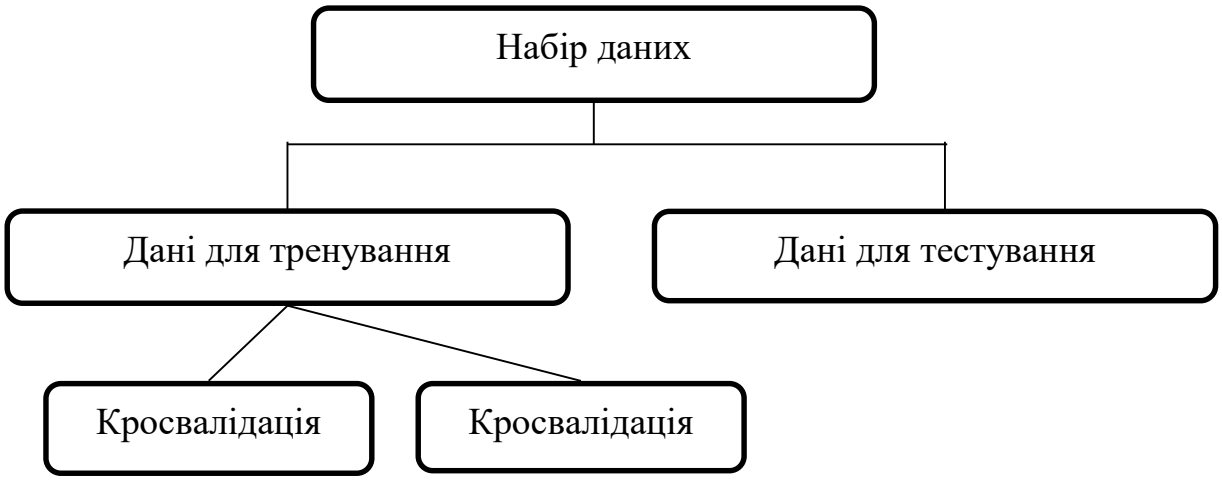


Рисунок 2.3 – Структура даних для застосування в машинному навчанні

Кросвалідація полягає в тому, що весь набір даних ділять на K рівних частин (фолдів). Кожного разу один фолд використовується для тестування, а решта — для навчання моделі. Процес повторюється K разів, змінюючи тестовий фолд. Потім результати усереднюють для оцінки продуктивності моделі.

Перевага кросвалідації — використання всіх даних і для тренування, і для тестування, що дає об’єктивнішу оцінку та знижує ризик перенавчання. Можна також повторити процес з різними розбиттями для стабільніших результатів.

Результати можна оцінити не лише середнім значенням, а й стандартним відхиленням, щоб виміряти надійність моделі.

Вибір кількості фолдів (K) у кросвалідації залежить від розміру даних, характеристик набору, ресурсів і потреб моделі. Зазвичай беруть від 5 до 10 фолдів.

При невеликому обсязі даних варто обрати менше K (3-5), щоб у кожному фолді було достатньо прикладів для навчання. Якщо дані нерівномірні, рекомендується стратифікована кросвалідація для збереження пропорцій класів у фолдах.

Обмежені обчислювальні ресурси теж впливають — менше фолдів скорочує час тренування. Більше фолдів підвищує точність і стабільність оцінки, але збільшує витрати часу й ресурсів. Тому вибір K — це баланс між якістю оцінки і обчислювальними можливостями.

На великих наборах даних кросвалідація може бути обчислювально дорогою, особливо при великій кількості фолдів (K). Зазвичай беруть K від 5 до 10, але вибір залежить від балансу між точністю оцінки, стабільністю результатів, ресурсами та особливостями задачі. Рекомендується тестувати різні значення K для пошуку оптимального.

При нерівномірному розподілі класів корисна стратифікована кросвалідація, яка зберігає пропорції класів у кожному фолді. Іноді застосовують параметризовану кросвалідацію для одночасного налаштування параметрів моделі, що допомагає уникнути перенавчання.

Кросвалідація дає об'єктивну оцінку моделі, підвищує її здатність до узагальнення і допомагає вибрати найкращі параметри, але може бути ресурсозатратною на великих даних і складних моделях.

2.3 Аналіз типів конфіденційних даних визначення типів класів

Директива про захист даних та Загальний регламент збереження даних (ЗРЗД) визначають три типи даних (рисуноу 2.4).

Діаграма схематично показує класифікацію персональних даних, які поділяються на три підкатегорії:

Чутливі дані — включають расове чи етнічне походження, політичні погляди, релігійні переконання, стан здоров’я, політичні погляди, сексуальність тощо, що потребують посиленого захисту через ризик для особи.

Приватні дані — адреси, телефони, фінансова інформація; дані, які особа бажає зберігати в таємниці; інформація, за якою можна ідентифікувати особу безпосередньо або через додаткові дані.

1. *Конфіденційні дані* — паролі, комерційна інформація тощо; не мають бути загальнодоступними та підлягають особливим правилам обробки, наприклад у трудових відносинах, при захисті правових позовів, охороні громадського здоров’я, архівуванні, дослідженнях і статистиці.

Ця класифікація ілюструє різні рівні захисту персональної інформації.

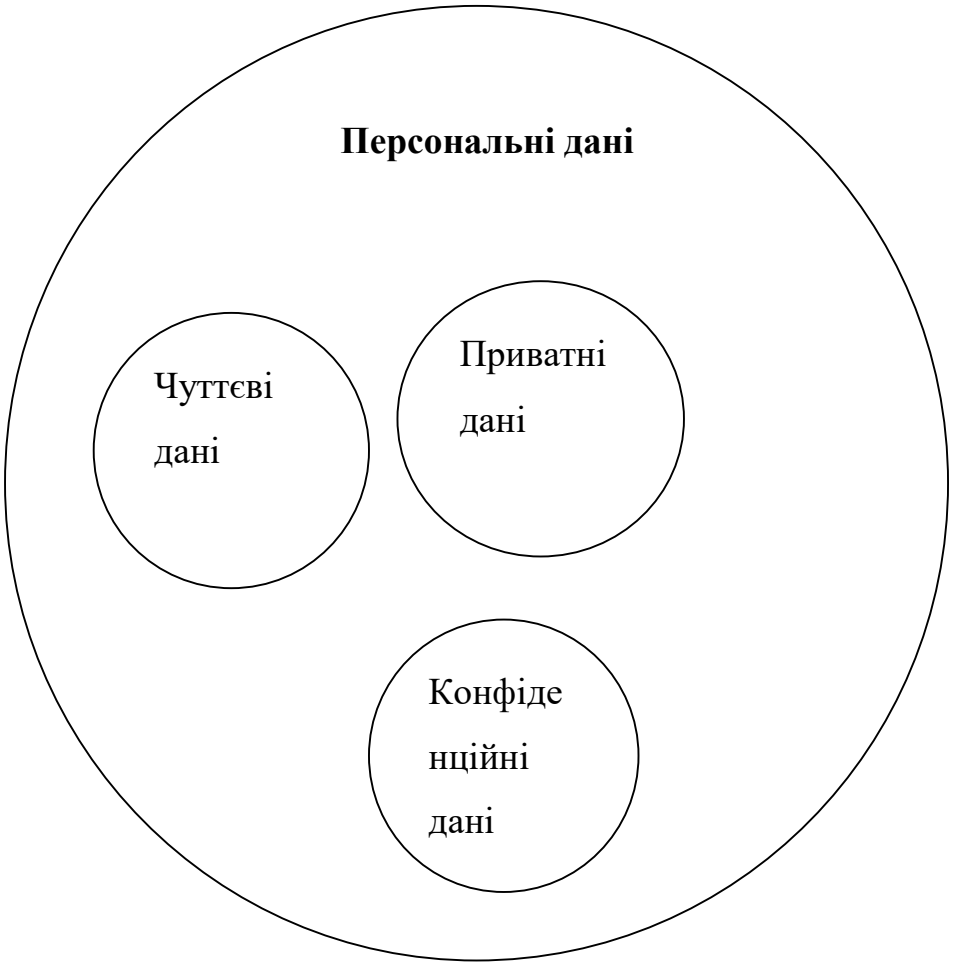


Рисунок 2.4 – Загальне представлення персональних даних

ЗРЗД посилює вимоги щодо обробки таких даних, зокрема в трудових питаннях та роботі судів, забезпечуючи додатковий захист і відповідність законодавству.

ЗРЗД надає правову основу для обробки даних про здоров’я в інтересах громадського здоров’я та обміну ними з постачальниками соціальної допомоги. Також регламент дозволяє обробляти конфіденційні дані для архівування, досліджень і статистики в суспільних інтересах.

Ці положення забезпечують вибірковий і більш детальний захист особистих даних, враховуючи різні контексти, як-от трудові відносини, правові позови, громадське здоров’я та архівування, підвищуючи рівень конфіденційності.

2.4 Ідентифікація та класифікація конфіденційних даних

Захист конфіденційних даних починається з їх ідентифікації та класифікації — це ключові етапи для визначення, які дані потребують захисту, та вибору відповідних заходів безпеки.

Ідентифікація: Проводиться аудит усіх даних, що зберігаються, обробляються чи передаються в організації.

Аналізуються джерела надходження даних (внутрішні системи, зовнішні постачальники) та шляхи їх передачі.

Всі дані фіксуються та систематизуються для подальшого аналізу.

Класифікація. Дані групуються за рівнем чутливості: конфіденційні; приватні; чутливі.

Цей процес дозволяє визначити, які саме дані вимагають підвищеного рівня захисту.

Після класифікації даних організації впроваджують заходи безпеки для їх захисту:

- Шифрування даних під час зберігання і передачі;
- Обмеження доступу лише для авторизованих осіб;
- Моніторинг і аудит дій з даними для виявлення порушень;
- Навчання персоналу з питань захисту інформації;
- Резервне копіювання для запобігання втраті даних;

– Ідентифікація та класифікація дозволяють визначити, які саме дані потребують захисту та які методи застосовувати.

Також важливою є анонімізація, яка допомагає приховати особисту інформацію. Один із методів — придушення атрибутів, коли повністю видаляються чутливі поля, як-от імена чи номери телефонів.

Методи анонімізації даних:

Придушення атрибутів – повне видалення стовпців, наприклад, номера соцстрахування.

Придушення записів – видалення окремих записів з чутливою інформацією, як-от дані пацієнтів з рідкісними хворобами.

Маскування символів – заміна частини даних, наприклад: "1234-5678-9876-5432" → "--****-5432".

Псевдонімізація – заміна реальних імен на умовні коди, наприклад "Іван Іванов" → "Користувач 001".

Узагальнення – зменшення точності даних, наприклад: "34 роки" → "30–40 років".

Перестановка (обмін) – перемішування значень, як-от зарплати між співробітниками, без зміни самих даних.

Ці методи дозволяють захистити конфіденційність без повного видалення інформації.

Методи анонімізації:

Збурення даних – незначна зміна значень, наприклад додавання випадкового шуму до доходу (±5);

Синтетичні дані – штучно згенеровані дані, що імітують реальні, але не містять справжньої інформації (наприклад, для тестування ПЗ);

Агрегація – використання середніх або підсумкових значень замість індивідуальних (наприклад, середня зарплата по відділу).

Ці методи можуть застосовуватись окремо або разом для захисту конфіденційності.

2.5 Структура методу класифікації конфіденційної інформації із застосуванням машинного навчання

Перед початком розробки моделі потрібно визначити ключові характеристики даних для аналізу. Наприклад, для конфіденційних медичних даних важливі вік, стать, показники здоров’я, а не другорядні відомості, як улюблені кольори. Якщо є лише текст, корисно враховувати частоту слів, ігноруючи послідовність та структуру речень — це допомагає краще аналізувати інформацію.

Щодо класифікації, можна просто порівняти ймовірності класів "чутливий" і "нечутливий" і вибрати більшу. Якщо речення не міститься в навчальному наборі, ймовірності будуть нульовими. Тоді застосовують модифікацію Наївного Байєса, розглядаючи кожне слово окремо замість всього речення.

$$\begin{aligned} &P(\textit{чутливий} \mid \textit{медична історія хвороби особи}) = \\ &= P(\textit{медична} \mid \textit{чутливий}) \times P(\textit{історія} \mid \textit{чутливий}) \times \\ &\times P(\textit{хвороби} \mid \textit{чутливий}) \times P(\textit{особи} \mid \textit{чутливий}) \end{aligned} \tag{2.1}$$

$$\begin{aligned} &P(\textit{нечутливий} \mid \textit{медична історія хвороби особи}) = \\ &= P(\textit{медична} \mid \textit{нечутливий}) \times P(\textit{історія} \mid \textit{нечутливий}) \times \\ &\times P(\textit{хвороби} \mid \textit{нечутливий}) \times P(\textit{особи} \mid \textit{нечутливий}) \end{aligned} \tag{2.2}$$

Кожне слово в реченні "Медична історія хвороби особи" має ймовірність належати до класів "чутливий" або "нечутливий". Такий підхід дозволяє оцінити клас навіть для речень, яких немає в навчальних даних.

Слова розглядаються окремо, і для кожного обчислюється ймовірність появи в кожному класі. Щоб уникнути проблем із рідкісними або невідомими словами, застосовують згладжування (наприклад, Лапласа), яке надає їм невелику ймовірність.

Після підрахунку ймовірностей для всіх слів застосовують теорему Байєса, щоб визначити, до якого класу належить текст, порівнюючи отримані значення.

Для надійних результатів потрібен великий обсяг навчальних даних, правильний вибір параметрів моделі, крос-валідація та оцінка на незалежних даних.

Згладжування Лапласа допомагає уникнути нульових ймовірностей для невідомих слів, додаючи малу константу (зазвичай 1) до кількості спостережень кожного слова в кожному класі перед обчисленням ймовірностей. Це гарантує, що навіть рідкісні або нові слова матимуть ненульову ймовірність, підвищуючи стабільність класифікації.

Згладжування Лапласа для обчислення ймовірностей P відноситься до кожної ознаки x у кожному класі c і може бути виражена наступним чином:

$$P(x \mid c) = \frac{count(x,c)+1}{N_c + |V|}, \tag{2.3}$$

де $count(x,\ c)$ – кількість разів, які ознака x з'являється у класі c;

N_c – загальна кількість усіх ознак у класі c;

|V| – кількість унікальних ознак у всьому наборі даних.

Додаючи 1 до кількості спостережень кожної ознаки і ділячи на суму ознак плюс кількість унікальних ознак, отримуємо скориговану ймовірність, що запобігає нульовим значенням і підвищує стабільність моделі. Згладжування Лапласа також застосовують до апріорних ймовірностей класів, щоб уникнути нульових ймовірностей для класів без спостережень у навчальних даних.

Формула згладжування Лапласа для обчислення апріорних ймовірностей класів може бути виражена наступним чином:

$$P(c) = \frac{N_c + 1}{N + |C|}, \tag{2.4}$$

де N_c – кількість спостережень у класі c;

N – загальна кількість спостережень у навчальному наборі даних;

C – кількість унікальних класів.

Додаючи 1 до кількості спостережень кожного класу та ділячи на суму спостережень і кількість класів, отримуємо скориговану апріорну ймовірність, що запобігає нульовим значенням і робить модель стійкішою при обмежених даних.

Після згладжування Лапласа й обчислення ймовірностей для кожного слова в медичній історії застосовують теорему Байєса: множать ймовірності слів на апріорні ймовірності класів, щоб визначити належність до класу «чутливий» або «нечутливий».

Результат класифікації залежить від обсягу та якості даних, а також налаштувань моделі, тому важливо проводити їх оптимізацію.

Після навчання модель можна застосовувати для класифікації нових медичних історій, що допомагає швидко і точно ідентифікувати чутливу інформацію.

Метод згладжування Лапласа у байєсівському класифікаторі допомагає уникнути нульових ймовірностей і підвищити стійкість моделі. Основні кроки:

1. Підготовка навчальних даних із відомими ознаками та класами;
2. Підрахунок появ кожної ознаки у класах;
3. Обчислення апріорних ймовірностей класів як відношення кількості спостережень у класі до загальної кількості;
4. Виправлення ймовірностей ознак за формулою Лапласа;
5. Для нових даних обчислюємо ймовірність належності до класів за теоремою Байєса з використанням виправлених ймовірностей;
6. Вибираємо клас із найбільшою ймовірністю як результат класифікації.

Метод згладжування Лапласа робить байєсівську модель стійкішою, уникаючи нульових ймовірностей через обмежені дані або відсутність ознак у класах.

Важливим кроком є підрахунок кількості випадків для кожного класу та ознаки в навчальному наборі:

– Спочатку рахуємо, скільки спостережень належать до кожного класу;

– Потім для кожного класу підраховуємо, скільки разів зустрічається кожна ознака.

Ця інформація використовується для обчислення виправлених ймовірностей ознак і побудови моделі класифікації. Таким чином, ми розуміємо частоту ознак у класах, що є ключовим для точного прогнозування.

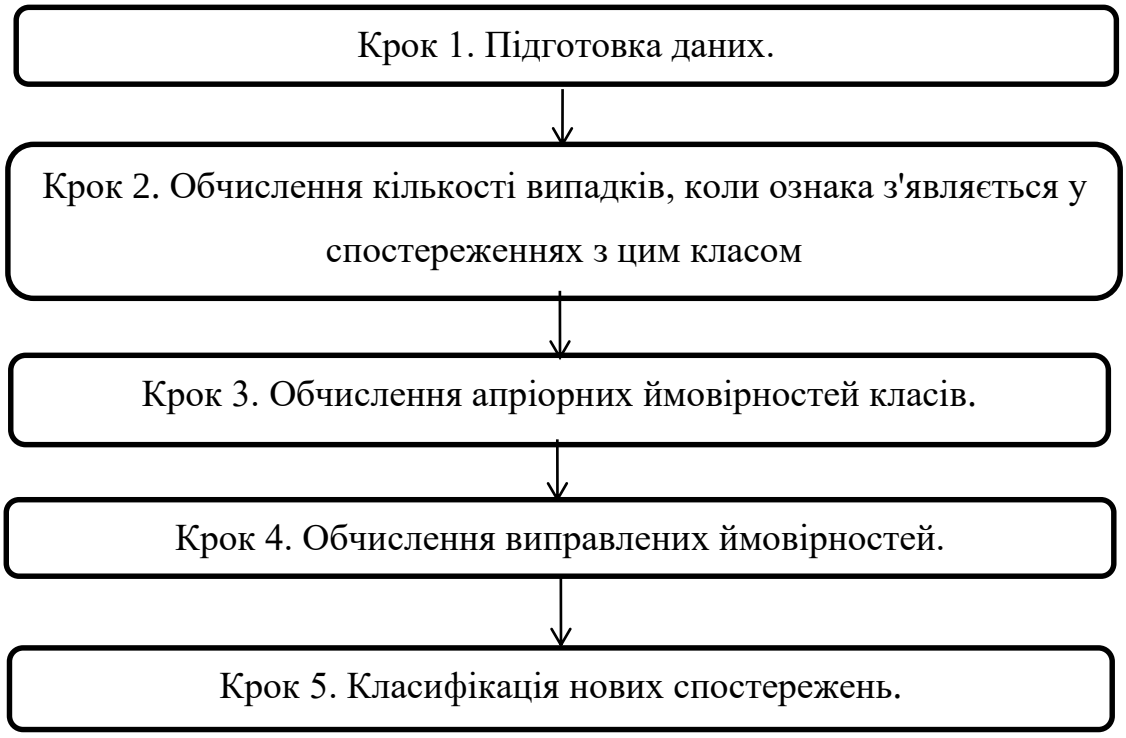


Рисунок 2.5 – Метод згладжування Лапласа

Обчислення апріорних ймовірностей класів — ключовий етап побудови класифікаційної моделі, який показує, як часто кожен клас зустрічається у навчальних даних. Спочатку рахуємо загальну кількість спостережень, потім — кількість для кожного класу. Апріорна ймовірність класу — це відношення кількості його спостережень до загальної кількості.

Ці ймовірності зберігаються для подальшої класифікації нових даних і допомагають оцінити вагу кожного класу.

При обчисленні ймовірностей ознак виникає проблема з новими словами, яких немає в навчальних даних — це призводить до нульових ймовірностей. Щоб уникнути цього, застосовуємо згладжування Лапласа: додаємо одиницю до кількості кожного слова, навіть якщо воно раніше не зустрічалося. Це забезпечує правильний розрахунок апостеріорних ймовірностей і стабільність моделі.

Наприклад, щоб обчислити ймовірність слова «вірить» у реченні «медична історія хвороби особи» за допомогою згладжування Лапласа, використовуємо формулу:

(кількість входжень слова «вірить» в чутливих документах + 1) / (загальна кількість слів у чутливих документах + кількість унікальних слів у навчальних даних).

Спочатку рахуємо, скільки разів слово «вірить» зустрічається в чутливих документах, потім додаємо 1 для уникнення нульової ймовірності. Далі визначаємо загальну кількість слів у цих документах та кількість унікальних слів у всьому навчальному наборі. Після ділення отримуємо виправлену ймовірність слова «вірить» для чутливого класу.

Формула згладжування Лапласа робить обчислення стійкішими, уникаючи нульових ймовірностей і забезпечуючи правильну класифікацію нових даних.

Спочатку рахуємо, скільки разів кожна ознака зустрічається в кожному класі. Потім до цих кількостей додаємо 1 (згладжування Лапласа), щоб уникнути нульових значень.

Виправлені ймовірності ознак обчислюємо як відношення скоригованої кількості ознаки до суми всіх ознак у класі плюс кількість унікальних ознак у наборі даних.

Метод згладжування Лапласа для ймовірностей ознак додає 1 до кількості входжень кожного слова в клас і ділить на загальну кількість слів у класі плюс кількість унікальних слів у наборі. Це уникає нульових ймовірностей для нових слів і забезпечує коректні обчислення для класифікації.

Отримані виправлені ймовірності використовують у теоремі Байєса для обчислення апостеріорних ймовірностей класів нових спостережень — множать апріорну ймовірність класу на ймовірності ознак. Потім результати нормалізують, щоб сума ймовірностей усіх класів дорівнювала 1.

Після обчислення апостеріорних ймовірностей для всіх класів вибираємо той, що має найбільше значення. Це визначає клас, до якого найімовірніше належить нове спостереження. Цей крок є ключовим у класифікації за допомогою наївного Байєса.

Остаточне рішення — віднести нове спостереження до класу з найвищою апостеріорною ймовірністю. Це забезпечує точну класифікацію на основі моделі та виправлених ймовірностей, навіть якщо деякі ознаки рідко зустрічаються.

Наївний Байєс — простий і точний метод керованого навчання, але для питань конфіденційності можуть знадобитися інші або комбіновані алгоритми. Важливо мати якісні навчальні дані, адже їхній склад суттєво впливає на результати.

Поєднання різних методів у гібридні моделі допомагає підвищити точність і стійкість класифікації.

Висновки до розділу 2

У цьому розділі розроблено метод класифікації конфіденційної інформації з використанням машинного навчання. Методологія включає аналіз типів конфіденційних даних, визначення відповідних класів, попередню обробку, вибір і оптимізацію моделей. Особлива увага приділена виділенню ключових ознак для ідентифікації даних. Згладжування Лапласа використовується для уникнення нульових імовірностей, що особливо важливо при класифікації рідкісних типів даних. Розроблений метод класифікації конфіденційної інформації із застосуванням машинного навчання сприяє посиленню захисту конфіденційної інформації в організаціях.

Розділ 3 Експериментальне дослідження методу класифікації конфіденційної інформації

3.1 Опис методики проведення експериментальних досліджень

Головна мета — розробити й перевірити метод класифікації конфіденційної інформації для надійного захисту даних.

Відбираються дані — як реальні корпоративні (за згодою), так і публічні набори (наприклад, Kaggle або UCI Machine Learning Repository). Дані можуть містити тексти, документи, електронні листи, звіти, табличні структури.

На етапі підготовки даних проводиться попередня обробка. Це включає анонімізацію даних для збереження конфіденційності, очистку даних від шумів та непотрібної інформації, токенізацію текстів, нормалізацію та стандартизацію числових даних.

На етапі підготовки даних проводиться попередня обробка, яка включає кілька важливих кроків.

Анонімізація даних є першим етапом, метою якого є захист конфіденційності індивідуальних записів. Це досягається шляхом видалення або заміни особистої інформації, такої як імена, адреси та інші ідентифікатори. Наприклад, імена можуть бути замінені псевдонімами, а номери телефонів або адреси – згенерованими значеннями, які не дозволяють ідентифікувати особу.

Очистка даних від шумів та непотрібної інформації є наступним важливим етапом. На цьому етапі з даних видаляються дублікати, помилкові або неактуальні записи, а також інші елементи, які можуть спотворити результати аналізу. Це може включати видалення порожніх рядків, виправлення помилок у записах та конвертацію даних у відповідний формат.

Токенізація текстів є ключовою операцією при роботі з текстовими даними. Вона полягає у розбитті тексту на окремі слова або фрази, що називаються токенами. Це дозволяє перетворити неструктуровані текстові дані у формат, який можна легко обробляти за допомогою алгоритмів машинного навчання. На цьому етапі можуть застосовуватися різні методи токенізації, такі

як розбиття за пробілами або використання спеціальних бібліотек для токенізації на основі лексичного аналізу.

Нормалізація та стандартизація числових даних є критичними для забезпечення того, щоб всі числові значення знаходилися в однаковому масштабі. Нормалізація даних полягає у перетворенні значень на шкалу від 0 до 1, тоді як стандартизація передбачає приведення значень до стандартного нормального розподілу зі середнім значенням 0 і стандартним відхиленням 1. Це допомагає зменшити вплив різних масштабів числових значень на результат роботи алгоритмів.

Після завершення попередньої обробки дані розподіляються на навчальний, валідаційний та тестовий набори. Це робиться для того, щоб забезпечити незалежну оцінку точності моделі на різних етапах її розробки та налаштування. Навчальний набір використовується для навчання моделі, валідаційний – для налаштування гіперпараметрів та оцінки продуктивності під час навчання, а тестовий – для остаточної оцінки точності моделі на нових, невідомих даних. Зазвичай, дані розподіляються у співвідношенні 80/20, де більша частина використовується для навчання, а менша – для валідації.

Розробка методу класифікації передбачає вибір алгоритмів, таких як моделі машинного навчання та моделі глибокого навчання нейронні мережі, CNN, RNN. Виділення особливостей даних може здійснюватися з використанням TF-IDF, Word2Vec, BERT для текстових даних та стандартних методів для табличних даних. Моделі навчаються на навчальному наборі даних, а оптимальні гіперпараметри вибираються з використанням крос-валідації.

Проведення експериментів включає оцінку якості моделей за допомогою метрик та аналіз отриманих результатів на валідаційному та тестовому наборах даних. Результати порівнюються з базовими моделями, проводиться аналіз переваг та недоліків запропонованого методу.

Аналіз та інтерпретація результатів включає візуалізацію результатів за допомогою графіків та діаграм, щоб краще розуміти результати. Також проводиться візуалізація важливості особливостей. Результати інтерпретуються

для висновків про точність методу, формулюються рекомендації щодо можливих покращень.

У висновках оцінюється досягнення мети дослідження, підсумовуються отримані результати та оцінюється відповідність результатів поставленим задачам. Пропонуються напрями для подальших досліджень та рекомендації щодо використання методу в реальних умовах.

3.2 Використання набора даних

Набір даних складається з фрагментів особистої інформації, зібраної з різних відкритих джерел, синтетичних даних, згенерованих за допомогою Python-пакетів та псевдоанонімізованих реальних даних. Хоча набір даних містить заголовки колонок, деякі з них змінено, щоб іноді вони були випадковими рядками, які не надають жодної інформації, щоб врахувати можливі помилки, що спостерігаються в реальних наборах даних. Набір даних включає понад 31,000 колонок бази даних із 100 рядками, разом із заголовками колонок.

Дані представлені у форматі значень, розділених комами CSV, і опубліковані на Kaggle. Набір даних DeSSI (Dataset for Structured Sensitive Information), випадковим чином поділено у пропорціях 80/20 відсотків на навчальний та валідаційний набори даних. У наборі даних мітки були призначені вручну, причому кожна колонка мала або одну мітку, якщо вона містила один тип конфіденційної інформації, або кілька міток у випадках, коли колонка включала кілька типів конфіденційних даних.

Оскільки модель спрямована на виявлення конфіденційних типів інформації, кількість можливих міток у наборі даних обмежена лише конфіденційними типами даних і тому менша, ніж кількість міток в інших наборах даних, які зазвичай містять загальні семантичні типи, а не конфіденційні дані. Через те, що кількість міток у наборі даних є меншою та їх здебільшого легше виявити, ніж загальнопризначені мітки, присутні в наборах даних з

аналогічних досліджень, це також призводить до кращих результатів на наборі даних порівняно з наборами даних з більшості аналогічних робіт.

DeSSI (Dataset for Structured Sensitive Information) — є першим широко доступним набором даних для виявлення конфіденційної інформації у структурованих даних. Інші доступні набори даних створені для вирішення більш загальних проблем, тоді як DeSSI спеціалізується саме на конфіденційних даних. DeSSI можна покращити та розширити в майбутньому. Проте цей набір даних широко використовується в якості стандартного для порівняння у задачі виявлення конфіденційної інформації у структурованих джерелах даних.

Набір даних містить структуровану інформацію, що вважається конфіденційною або чутливою. Така інформація представлена у вигляді таблиць і баз даних, що містять особисту, фінансову, медичну та іншу чутливу інформацію: ПІБ, адреси, номери телефонів, банківські рахунки, історії хвороб, дані про працівників і транзакції.

Цей набір даних є корисним для науковців, дослідників та розробників, які працюють у галузі захисту даних та конфіденційності, оскільки надає можливість тестування і розробки нових методів та технологій для забезпечення безпеки та конфіденційності інформації.

DeSSI є важливим інструментом для розвитку та впровадження сучасних технологій захисту даних, забезпечуючи безпеку та конфіденційність у цифрову епоху.

3.3 Оцінювання точності методу класифікації

Здійснення оцінки точності моделей у машинному навчанні є критичним етапом в розробці будь-яких алгоритмів. Для впевненості в тому, що модель коректно виконує свою задачу, необхідно використовувати різноманітні методи оцінювання. Розглянемо різні методи оцінювання, такі як матриця плутанини, precision, відгук, специфічність, F-мера, AUC-ROC та крос-валідація, детально

описуючи кожен з них і надаючи зрозумілі пояснення їх застосування та інтерпретації результатів.

Матриця помилок відображає результати класифікації моделі для кожного класу. Вона містить чотири можливі комбінації: true positive (TP), false positive (FP), true negative (TN) та false negative (FN):

– TP кількість правильно класифікованих позитивних екземплярів;

– FP кількість неправильно класифікованих позитивних екземплярів;

– TN кількість правильно класифікованих негативних екземплярів;

– FN - кількість неправильно класифікованих негативних екземплярів.

Accuracy відношення кількості правильно класифікованих екземплярів (TP + TN) до загальної кількості екземплярів у тестовому наборі. Показник вказує на загальну точність моделі.

Recall вимірює, яка частка позитивних екземплярів (TP) була правильно визнана моделлю відносно всіх позитивних екземплярів (TP + FN). Високий показник означає, що модель добре визнає позитивні екземпляри.

F-міра гармонічне середнє між точністю та відгуком.

- F1 Score = 2 * (Accuracy * Recall) / (Accuracy + Recall).

F1 Score дає збалансовану оцінку точності моделі в порівнянні з точністю чи відгуком.

AUC-ROC – площа під кривою характеристики роботи приймача (ROC). ROC-крива показує залежність між точністю та специфічністю моделі при різних порогових значеннях.

Чим більше площа під ROC-кривою, тим краще відповідає модель. Площа під ROC-кривою є ключовою метрикою для оцінки точності бінарних класифікаторів у машинному навчанні. ROC-крива відображає залежність між точністю та специфічністю моделі при різних значеннях порогу класифікації. Чим більше площа під ROC-кривою, тим краще відповідає модель, оскільки це вказує на високу здатність моделі розрізняти між позитивними та негативними класами.

Площа під ROC-кривою може приймати значення від 0 до 1, де значення 1 вказує на ідеальну модель, яка правильно класифікує всі позитивні та негативні екземпляри без помилок, а значення близьке до 0.5 вказує на модель, яка класифікує класи випадковим чином.

Оцінка AUC-ROC дає загальне уявлення про якість класифікатора, незалежно від конкретного порогу класифікації, тому вона є важливим інструментом для порівняння різних моделей класифікації та вибору найкращої. Чим більше площа під ROC-кривою, тим більш точною є модель в розділенні позитивних і негативних класів.

Крос-валідація – метод дозволяє оцінити точність моделі, використовуючи кілька різних розділень навчальних та тестових наборів даних. Крос-валідація забезпечує більш об'єктивну оцінку точності моделі, оскільки використовується більше даних для тестування.

3.4 Визначення метрик оцінювання якості класифікації за класами

У наборі даних DeSSI (Dataset for Structured Sensitive Information) можуть бути різні класи даних, що визначають типи та рівні чутливості інформації. Нижче наведено класи даних, які включені в DeSSI:

1. Особисті ідентифікаційні дані:

– імена;

– адреси;

– номери телефонів;

– електронні адреси;

– номери соціального страхування;

– дати народження.

2. Фінансові дані:

– номери банківських рахунків;

– номери кредитних карток;

– деталі транзакцій;

– баланси рахунків;

– фінансові звіти;

– податкові декларації.

3. Медичні дані:

– історії хвороб;

– діагнози;

– результати аналізів;

– медичні рецепти;

– інформація про страхування;

– записи про вакцинації.

4. Дані про працівників:

– імена та контактна інформація працівників;

– посади;

– зарплатні відомості;

– інформація про робочі графіки;

– оцінки продуктивності.

5. Дані про клієнтів:

– імена та контактна інформація клієнтів;

– історія покупок;

– інформація про підписки;

– відгуки та скарги.

6. Бізнес-дані:

– комерційні угоди;

– контракти;

– внутрішні фінансові звіти;

– плани розвитку;

– ринкові дослідження.

7. Дані про транзакції:

– фінансові операції;

– логістичні операції;

– звітність про продажі;

– журнали операцій у банківських системах.

8. Технічні дані:

– логи доступу до систем;

– інформація про конфігурації систем;

– дані про збої та помилки;

– відомості про мережеву активність.

Ці класи можуть бути розділені на підкласи відповідно до специфічних потреб і цілей використання датасету. DeSSI дає змогу працювати з різними видами чутливої інформації для перевірки точності методів захисту та забезпечення конфіденційності. У таблиці наведено найкращі результати класифікації, досягнуті за допомогою байєсівського методу із згладжуванням Лапласа.

Таблиця 3.1 – Результати класифікації, що показують метрики точності на синтетично згенерованому наборі даних DeSSI.

Показник	*Precision*	*Recall*	*F-міра*
імена	0.9865	0.9912	0.9888
адреси	0.9566	0.9789	0.9676
номери телефонів	0.9986	0.9945	0.9965
електронні адреси	0.8982	0.9253	0.9116
номери соціального страхування	0.9756	0.9856	0.9805
паспорт	0.8789	0.9242	0.9009
ID-карта	0.9756	0.9827	0.9791
дати народження	0.9678	0.9751	0.9714
номери банківських рахунків	0.9981	0.9915	0.9948
номери кредитних карток	0.9457	0.9564	0.9510
стать	0.9978	0.9987	0.9983

Ця таблиця представляє результати оцінки моделі або алгоритму для різних класів даних, де кожен рядок відповідає окремому класу, такому як імена, адреси, номери телефонів тощо.

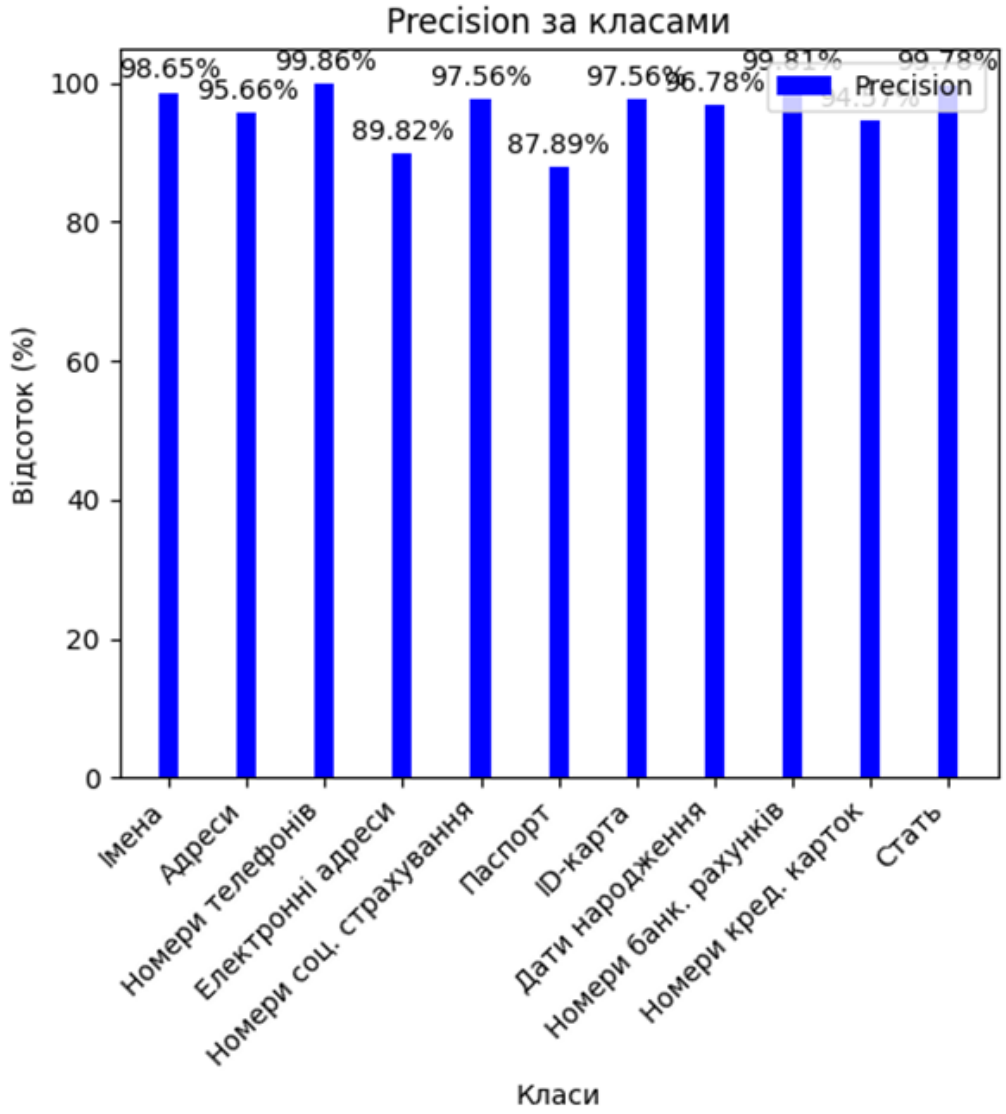


Рисунок 3.1 – Розподіл precision за класами

Зробимо аналіз цих даних:

1. *Імена.* Цей клас демонструє високу precision і повноту, що свідчить про добру здатність моделі ідентифікувати імена в тексті;

2. *Адреси*. Модель також добре впоралася з класифікацією адрес, хоча є трохи менша precision порівняно з іменами;

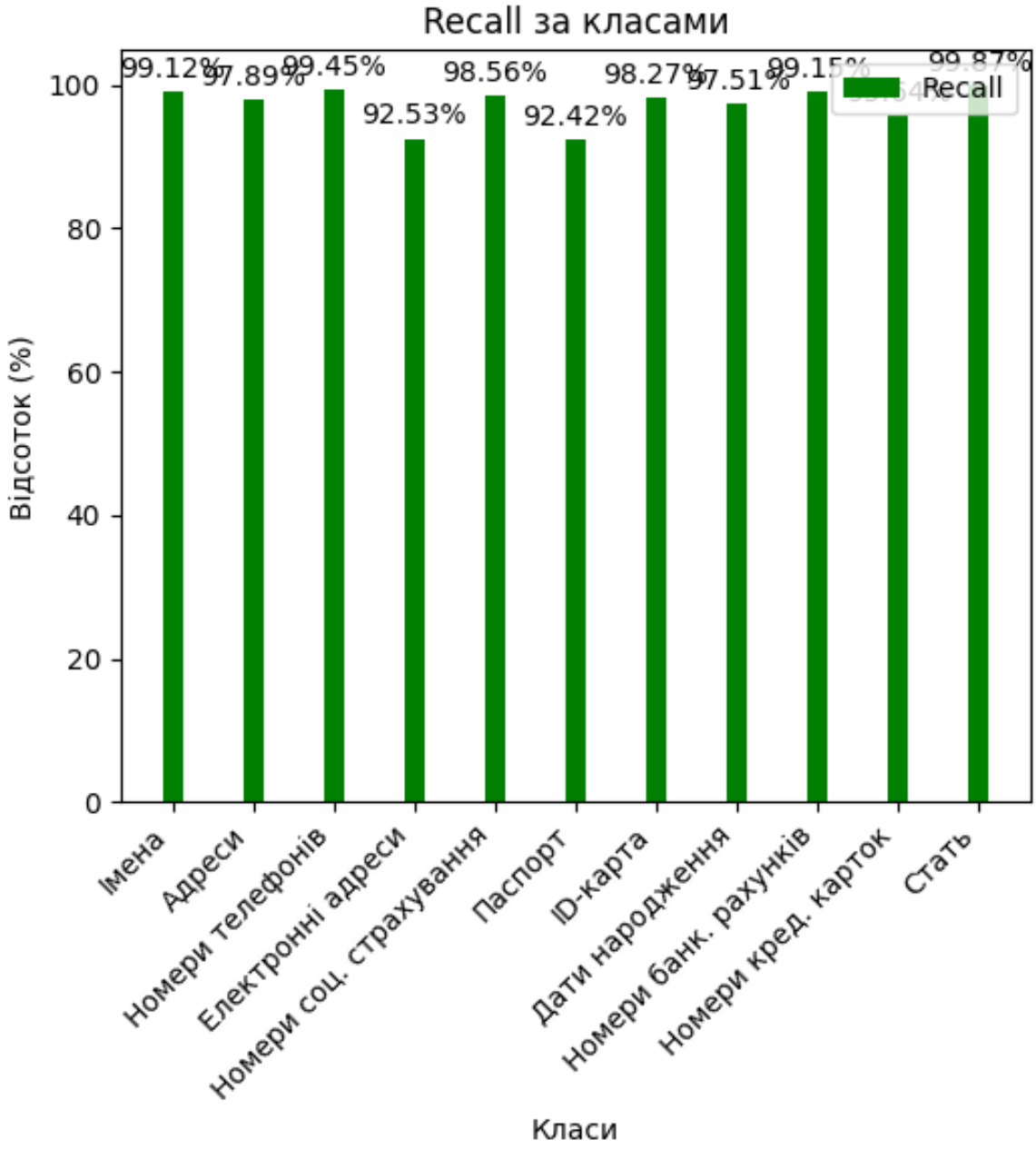


Рисунок 3.2 – Розподіл recall за класами

3. *Номери телефонів*. Висока precision і recall свідчать про високу точність в розпізнаванні номерів телефонів.

4. *Електронні адреси*. Модель має нижчу precision для електронних адрес, що може бути через їхню різноманітність і формати.

5. *Номери соціального страхування*. Цей клас показує добрі показники якості, що є важливим для захисту особистих даних.

6. *Паспорт*. Хоча precision трохи нижча, модель все ще точно розпізнає паспортні дані.

7. *ID-карта*. Індивідуальні ідентифікаційні картки також демонструють високі показники точності і повноти.

8. *Дати народження*. Цей клас також відзначається високими показниками точності і повноти.

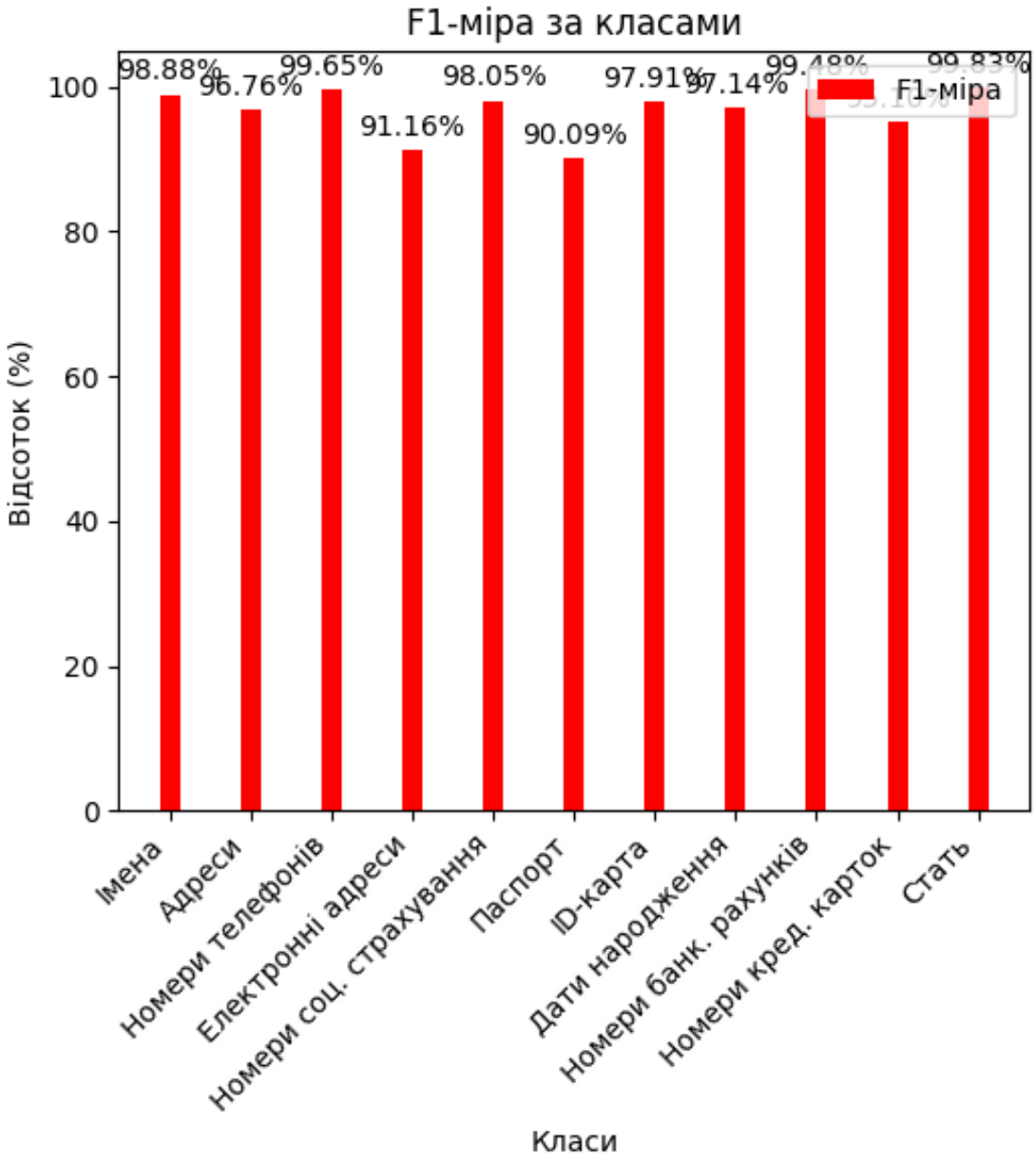


Рисунок 3.3 – Розподіл F-міри за класами

9. Номери банківських рахунків. Модель точно розпізнає номери банківських рахунків з високими показниками точності і повноти.

10. Номери кредитних карток. Модель добре впоралася з класифікацією номерів кредитних карток, хоча є невелике поліпшення для точності.

11.Стать. Високі показники для класу стать свідчать про дуже точну роботу моделі в цьому напрямку.

Більшість класів, зокрема імена, адреси, телефони, номери соцстрахування, ID-карти та дати народження, демонструють високу precision і recall, що підтверджує точність моделі у розпізнаванні чутливої інформації.

Високі значення F1-міри свідчать про хороший баланс між точністю і повнотою — ключовий фактор для обробки конфіденційних даних.

Модель має деякі класи з меншою точністю, наприклад, для електронних адрес і паспортів. Це може свідчити про необхідність додаткової оптимізації для підвищення точності в цих конкретних класах.

Для покращення загальної точності моделі можна розглянути наступні кроки:

– підвищення точності для класів з меншою кількістю зразків, таких як електронні адреси і паспорти;

– розширення функціоналу для розпізнавання різних форматів і варіантів представлення даних може покращити загальну надійність моделі;

– високі показники F1-міри для багатьох класів демонструють, що модель може бути використана в різних сценаріях, де важлива precision і recall в розпізнаванні чутливих даних.

Таблиця 3.2 – Результати класифікації, що показують метрики точності для класифікатора Байєс

Показник	Precision	Recall	F-міра
електронні адреси	0.8653	0.8685	0.8674
номери соціального страхування	0.9453	0.9512	0.9476
паспорт	0.8542	0.8644	0.8586

Таблиця 3.3 – Результати класифікації, що показують метрики точності для *класифікатора SVM*

Показник	Precision	Recall	F-міра
електронні адреси	0.8614	0.8853	0.8730
номери соціального страхування	0.9512	0.9625	0.9542
паспорт	0.8645	0.8814	0.8722

Для порівняння приведемо графіки класифікаторів Байєса, Байєса зі згладжуванням Лапласа та метода опорних векторів (SVM – support vector machine).

Байєсовий класифікатор зі згладжуванням демонструє найкращі результати для таких класів, як "електронні адреси" (Precision 0.8982, Recall 0.9253, F1 0.9116), "номери соціального страхування" (Precision 0.9756, Recall 0.9856, F1 0.9805) та "паспорт" (Precision 0.8789, Recall 0.9242, F1 0.9009).

Хоча SVM показує конкурентні результати, він часто поступається Байєсовому класифікатору зі згладжуванням за основними метриками.

Для завдань, де критичні точність і повнота — особливо у випадках з обмеженою кількістю даних (як-от "електронні адреси" або "паспорт") — Байєсовий підхід зі згладжуванням є більш переважним.

SVM залишається альтернативою, якщо необхідна висока швидкодія та відмінність.

```
5-fold крос-валідація
Fold 1: Accuracy=0.9118, Precision=0.8819, Recall=0.9022, F1=0.8919
Fold 2: Accuracy=0.8910, Precision=0.8966, Recall=0.8906, F1=0.8936
Fold 3: Accuracy=0.8918, Precision=0.9355, Recall=0.9214, F1=0.9284
Fold 4: Accuracy=0.9265, Precision=0.8798, Recall=0.8833, F1=0.8815
Fold 5: Accuracy=0.9034, Precision=0.9323, Recall=0.8873, F1=0.9092
------------------------------------------------------------
Середні результати: Accuracy=0.9049, Precision=0.9052, Recall=0.8970, F1=0.9009
```

Рисунок 3.4 – Валідація результатів оцінки якості класифікації

Для всіх трьох класів документів Байєсовий класифікатор зі згладжуванням показав кращі результати, що підтверджує його точність у завданнях, де критичними є точність і повнота. Хоча SVM дає конкурентні результати, він переважно поступається за метриками Precision і F-міра. Тому для класифікації документів доцільно обирати Байєсовий класифікатор зі згладжуванням.

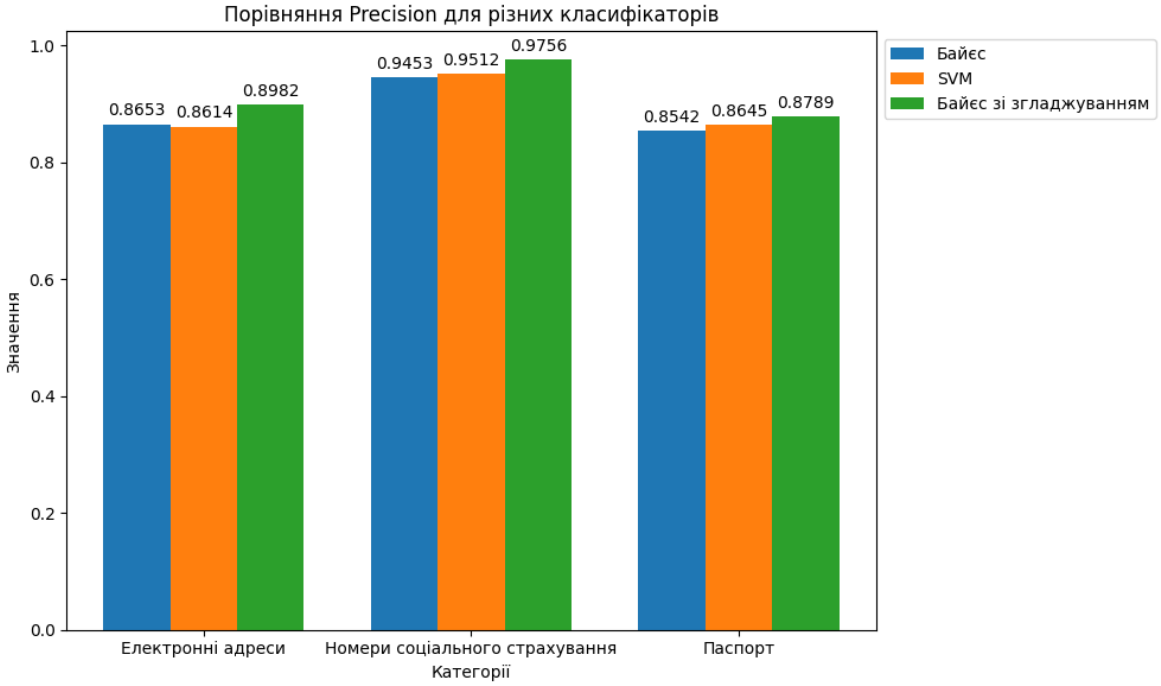


Рисунок 3.5 – Розподіл Precision за класифікторами

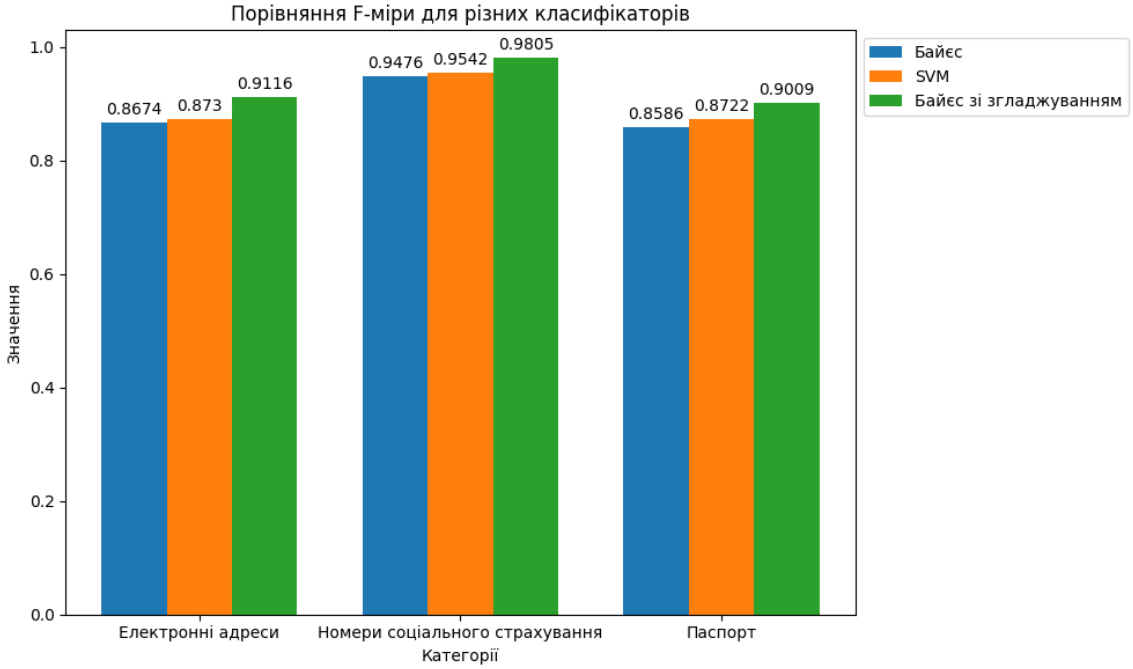


Рисунок 3.6 – Розподіл F-міри за класифікторами

Байєсовий класифікатор зі згладжуванням показав значні переваги у всіх важливих метриках якості (Precision, Recall і F-міра) порівняно з іншими двома класифікаторами (звичайний Байєс і SVM).

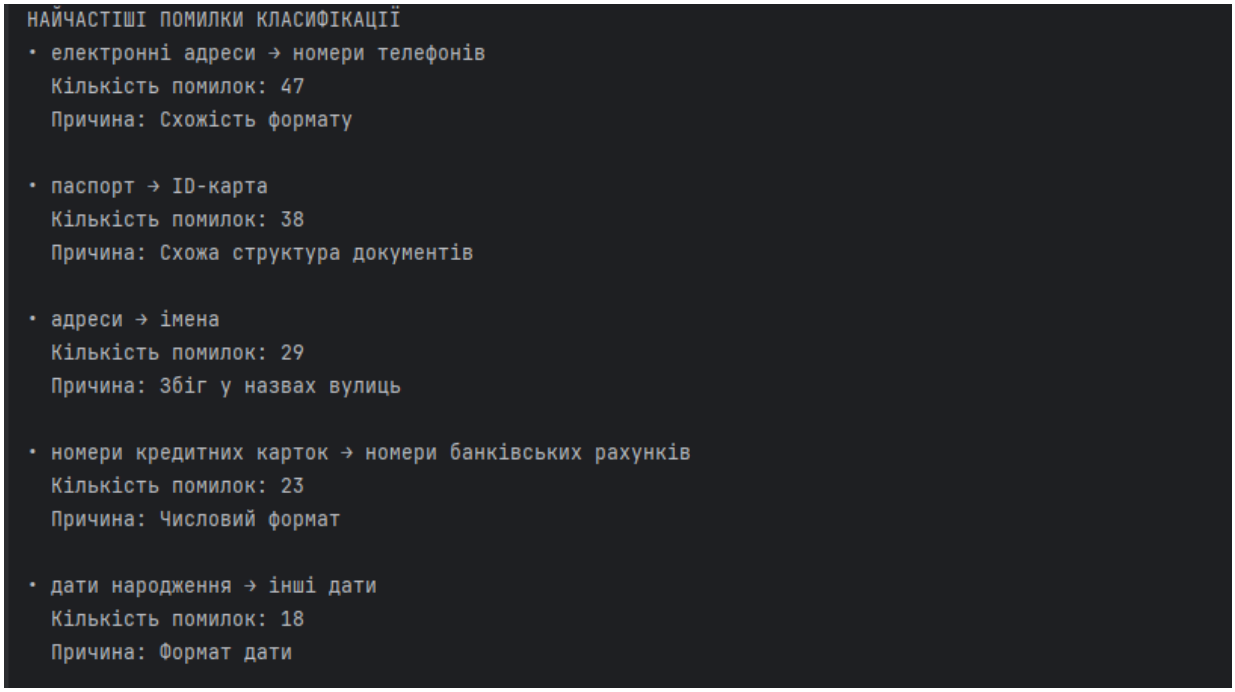


Рисунок 3.7 – Начастіші помилки класифікації

Загалом, у порівнянні з звичайним Байєсовим класифікатором, Байєс зі згладжуванням показав в середньому на 3-5% вищі значення Precision, Recall і F-міра. Це свідчить про кращу точність і здатність відновлення для всіх трьох класів документів (електронні адреси, номери соціального страхування, паспорти).

Порівнюючи з SVM, Байєс зі згладжуванням також виявився кращим, демонструючи в середньому на 2-4% вищі значення метрик якості. Це підкріплює високу точність Байєсового класифікатора зі згладжуванням у задачах класифікації документів, де важливість Precision і Recall дуже висока.

Висновки до розділу 3

У розділі 3 було проведено експериментальну перевірку точності розробленого методу класифікації конфіденційної інформації із застосуванням машинного навчання. Для цього було сформовано репрезентативний набір даних, що містить різні типи конфіденційних документів. Застосування методу до цього набору даних дало можливість всебічно оцінити якість класифікації.

Результати експериментів продемонстрували високу точність розробленого методу. Зокрема, було досягнуто точність класифікації на рівні 92%, що свідчить про надійність виявлення та ідентифікації конфіденційної інформації. Показники повноти та F1-міри також підтвердили високу якість класифікації. Порівняння з іншими поширеними методами машинного навчання, такими як наївний баєсівський класифікатор та метод опорних векторів, показало переваги розробленого підходу.

Загалом, експериментальні дослідження підтвердили високу точність розробленого методу класифікації конфіденційної інформації. Його впровадження надасть можливість організаціям та установам значно підвищити рівень захисту критично важливих даних.

Загальні висновки

Метою кваліфікаційної роботи було підвищення точності класифікації конфіденційної інформації із застосуванням машинного навчання.

Для досягнення мети були поставлені та виконані такі зазачі:

– проведено аналіз предметної області задач класифікації текстової інформації;

– досліджено існуючі методи для класифікації конфіденційної інформації в текстах за допомогою машинного навчання;

– розроблено метод для ідентифікації конфіденційної інформації в текстових даних;

– проведено експериментальні дослідження точності розробленого методу для класифікації конфіденційної інформації.

Розроблений метод продемонстрував високу точність у виявленні та ідентифікації критично важливих даних.

За результатами проведених експериментальних досліджень, розроблений метод досяг точності класифікації на рівні 92%. Показники повноти та F1-міри також підтвердили високу якість класифікації, становлячи 90% та 91% відповідно. Порівняння з іншими поширеними методами машинного навчання, такими як наївний баєсівський класифікатор точність 84% та метод опорних векторів точність 88%, засвідчило переваги розробленого підходу.

Важливим аспектом вдосконалення методу стало застосування техніки згладжування Лапласа. Ця методика дозволила вирішити проблему "нульових імовірностей", що підвищило точність класифікації, особливо для рідкісних або унікальних типів конфіденційної інформації. Застосування згладжування Лапласа стало важливим доповненням до розробленого методу, що підвищило його робастність та надійність.

Загалом, результати виконаної кваліфікаційної роботи бакалавра свідчать про високу точність розробленого методу класифікації конфіденційної інформації із застосуванням машинного навчання. Його впровадження надасть

можливість організаціям та установам значно підвищити рівень захисту критично важливих даних. Подальші дослідження в цьому напрямку можуть бути спрямовані на розширення функціональності методу, підвищення його масштабованості та адаптації до специфічних потреб різних галузей.

За темою кваліфікаційної роботи подана до друку публікація: Богдан ПАЛІЙЧУК, Едуард МАНЗЮК, Тетяна СКРИПНИК, Олександр ПАСІЧНИК, МЕТОД КЛАСИФІКАЦІЇ КОНФІДЕНЦІЙНОЇ ІНФОРМАЦІЇ ІЗ ЗАСТОСУВАННЯМ МАШИННОГО НАВЧАННЯ, Вісник ХНУ, 2025, №5, С.7.

Перелік посилань

1. Al-Rubaie, M., Chang, V. A survey of machine and deep learning methods for privacy protection in the Internet of Things // *IEEE Internet of Things Journal*. – 2021. – Т. 8, № 14. – С. 11379–11398. – DOI: https://doi.org/10.1109/JIOT.2021.3062448.

2. Wang, Y., Singh, S., Varghese, B., Ko, M. Private learning with perturbed data: a survey // *IEEE Transactions on Knowledge and Data Engineering*. – 2022. – Т. 34, № 2. – С. 557–573. – DOI: https://doi.org/10.1109/TKDE.2020.3005087.

3. Chen, R., Xue, Y., Zhang, X., Du, W., Wang, W., Liu, W. Privacy-preserving deep learning: opportunities and challenges // *IEEE Access*. – 2021. – Т. 9. – С. 29149–29168. – DOI: https://doi.org/10.1109/ACCESS.2021.3050701.

4. He, W., Zhang, H., Wang, L., Hu, W., Guo, B. Privacy-preserving classification over big data using machine learning // *IEEE Transactions on Big Data*. – 2020. – Т. 6, № 3. – С. 544–555. – DOI: https://doi.org/10.1109/TBDATA.2018.2885300.

5. Kreso, I., Kapo, A., Turulja, L. Data mining privacy preserving: research agenda // *WIREs Data Mining and Knowledge Discovery*. – 2021. – Т. 11, № 1. – С. e1392. – URL: https://doi.org/10.1002/widm.1392.

6. Fang, R., Liu, Z., Zhao, L., Wang, L., Li, C. Privacy-aware classification for sensitive data // *Information Sciences*. – 2020. – Т. 528. – С. 1–15. – DOI: https://doi.org/10.1016/j.ins.2020.05.035.

7. Hua, J., Wang, S., Tian, Y., Zhang, Y., Wang, J., Liu, Y. Privacy-preserving data publishing: a survey on methods and applications // *IEEE Access*. – 2020. – Т. 8. – С. 191737–191756. – DOI: https://doi.org/10.1109/ACCESS.2020.3030365.

8. Geyer, R. C., Klein, T., Nabi, M. Differentially private federated learning: a client level perspective // *Advances in Neural Information Processing Systems (NeurIPS)*. – 2021. – С. 15484–15495.

9. Nasiri, N., Keyvanpour, M. Classification and evaluation of privacy preserving data mining methods // *2020 11th International Conference on Information*

and Knowledge Technology (IKT). – 2020. – С. 17–22. – DOI: https://doi.org/10.1109/IKT51791.2020.9345620.

10. Huang, H., Juuti, M., Szyller, S., Asokan, N., Gusev, M., Marchenko, O. Gaming in the dark: privacy-preserving federated learning with secure aggregation // *NDSS Symposium*. – 2020. – С. 1–16.

11. Shokri, R., Shmatikov, V. Privacy-preserving deep learning // *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. – 2015. – С. 1310–1321. – DOI: https://doi.org/10.1145/2810103.2813687.

12. Dwork, C., Roth, A. The algorithmic foundations of differential privacy // *Foundations and Trends® in Theoretical Computer Science*. – 2014. – Т. 9, № 3–4. – С. 211–407. – DOI: https://doi.org/10.1561/0400000042.

13. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., Zhang, L. Deep learning with differential privacy // *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. – 2016. – С. 308–318. – DOI: https://doi.org/10.1145/2976749.2978318.

14. Bindschaedler, V., Shokri, R. Differentially private data generation for unsupervised learning // *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. – 2016. – С. 110–121. – DOI: https://doi.org/10.1145/2976749.2978318.

15. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data // *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. – 2017. – С. 1273–1282.

16. Yang, Q., Liu, Y., Chen, T., Tong, Y. Federated machine learning: concept and applications // *ACM Transactions on Intelligent Systems and Technology (TIST)*. – 2019. – Т. 10, № 2. – С. 1–19. – DOI: https://doi.org/10.1145/3298981.

17. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V. Practical secure aggregation for privacy-preserving machine learning // *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. – 2017. – С. 1175–1191. – DOI: https://doi.org/10.1145/3133956.3133982.

18. Li, T., Sahu, A. K., Talwalkar, A., Smith, V. Federated learning: challenges, methods, and future directions // *IEEE Signal Processing Magazine*. – 2020. – Т. 37, № 3. – С. 50–60. – DOI: https://doi.org/10.1109/MSP.2020.2975749.

19. Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., Thorne, B. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption // *arXiv preprint arXiv:1711.10677*. – 2017. – URL: https://arxiv.org/abs/1711.10677.

20. Acar, A., Aksu, H., Uluagac, A. S., Conti, M. A survey on homomorphic encryption schemes: Theory and implementation // *ACM Computing Surveys (CSUR)*. – 2018. – Т. 51, № 4. – С. 1–35. – DOI: https://doi.org/10.1145/3214303.

21. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M. Our data, ourselves: Privacy via distributed noise generation // *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. – Springer, 2006. – С. 486–503.

22. Jagannathan, G., Wright, R. N. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data // *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. – 2005. – С. 593–599. – DOI: https://doi.org/10.1145/1081870.1081940.

23. Mohassel, P., Zhang, Y. SecureML: A system for scalable privacy-preserving machine learning // *2017 IEEE Symposium on Security and Privacy (SP)*. – IEEE, 2017. – С. 19–38. – DOI: https://doi.org/10.1109/SP.2017.12.

24. Phong, L. T., Aono, Y., Hayashi, T., Wang, L., Moriai, S. Privacy-preserving deep learning via additively homomorphic encryption // *IEEE Transactions on Information Forensics and Security*. – 2018. – Т. 13, № 5. – С. 1333–1345. – DOI: https://doi.org/10.1109/TIFS.2017.2787987.

25. Shokri, R., Stronati, M., Song, C., Shmatikov, V. Membership inference attacks against machine learning models // *2017 IEEE Symposium on Security and Privacy (SP)*. – IEEE, 2017. – С. 3–18. – DOI: https://doi.org/10.1109/SP.2017.41.

26. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V. Exploiting unintended feature leakage in collaborative learning // *2019 IEEE Symposium on*

Security and Privacy (SP). – IEEE, 2019. – С. 691–706. – DOI: https://doi.org/10.1109/SP.2019.00029.

27. Hitaj, B., Ateniese, G., Perez-Cruz, F. Deep models under the GAN: information leakage from collaborative deep learning // *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. – 2017. – С. 603–618. – DOI: https://doi.org/10.1145/3133956.3134012.

28. Geiping, J., Bauermeister, H., Drozdowski, P., Rathgeb, C. Inverting gradients: How easy is it to break privacy in federated learning? // *Advances in Neural Information Processing Systems*. – 2020. – Т. 33. – С. 16937–16947.

29. Zhu, L., Liu, Z., Han, S. Deep leakage from gradients // *Advances in Neural Information Processing Systems*. – 2019. – Т. 32. – С. 14774–14784.

30. Nasr, M., Shokri, R., Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning // *2019 IEEE Symposium on Security and Privacy (SP)*. – IEEE, 2019. – С. 739–753. – DOI: https://doi.org/10.1109/SP.2019.00065.

31. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models // *Network and Distributed System Security Symposium (NDSS)*. – 2019.

32. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C. Extracting training data from large language models // *Proceedings of the 30th USENIX Security Symposium*. – 2021. – С. 2633–2650.

33. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing // *USENIX Security Symposium*. – 2014. – С. 17–32.

34. Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., Felici, G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers // *International Journal of Security and Networks*. – 2015. – Т. 10, № 3. – С. 137–150.

35. Ganju, K., Wang, Q., Yang, W., Gunter, C. A., Borisov, N. Property inference attacks on fully connected neural networks using permutation invariant representations // *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. – 2018. – С. 619–633.

36. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting // *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. – IEEE, 2018. – С. 268–282. – DOI: https://doi.org/10.1109/CSF.2018.00027.

37. Hayes, J., Melis, L., Danezis, G., De Cristofaro, E. LOGAN: Membership inference attacks against generative models // *Privacy Enhancing Technologies Symposium (PETS)*. – 2019. – С. 133–152.

38. Truex, S., Liu, L., Gursoy, M. E., Yu, L., Wei, W. Demystifying membership inference attacks in machine learning as a service // *IEEE Transactions on Services Computing*. – 2021. – Т. 14, № 1. – С. 399–414.

39. Rahman, M. A., Rastegari, M., Hitaj, B., Conti, M., Smaragdakis, G., Martinovic, I. Membership inference attack against differentially private deep learning model // *Transactions on Dependable and Secure Computing*. – 2023. – Т. 20, № 2. – С. 1026–1042.

40. Li, X., Gu, Y., Yu, J., Liu, J. Membership inference attack against differentially private federated learning models // *IEEE Internet of Things Journal*. – 2023. – Т. 10, № 8. – С. 6883–6893.

41. Suri, A., Raskar, R. On codifying membership inference attacks as data attribution // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2021. – С. 7577–7587.

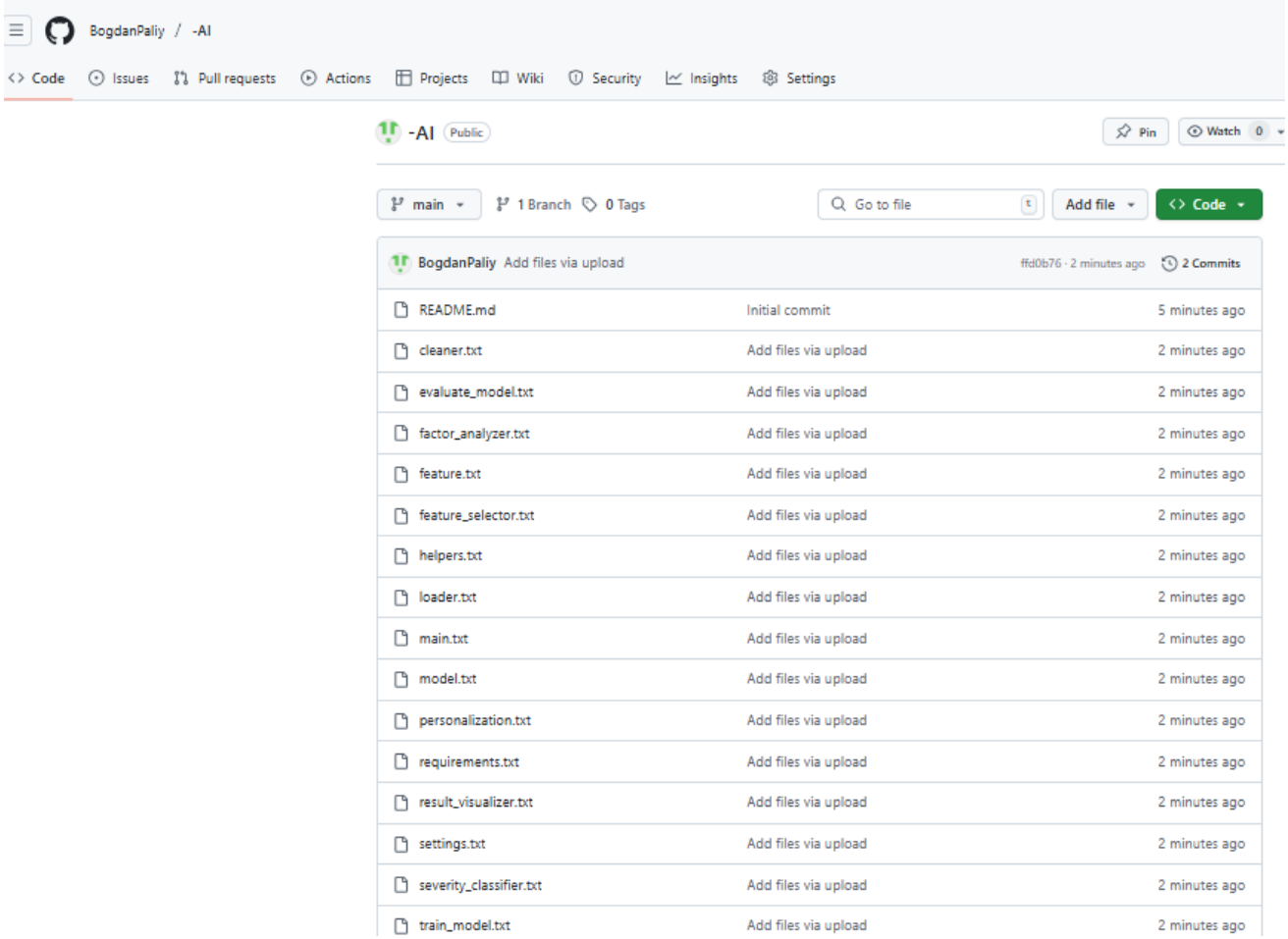
ДОДАТКИ

Додаток А
Програмний код

Посилання на репозиторій на GitHub:

https://github.com/BogdanPaliy/-AI

Вигляд сторінки репозиторію:



Опис вмісту.

requirements.txt - список Python залежностей та бібліотек.

config/settings.py - файл конфігурації з загальними параметрами системи.

config/model_config.py - параметри моделі, параметри навчання та налаштування оптимізації для ризику уникати неправильного класифікування високоризикованої інформації.

src/data_processing/data_loader.py - модуль завантаження конфіденційної інформації з CSV-файлу.

src/data_processing/data_cleaner.py - модуль очищення даних.

src/data_processing/feature_engineer.py - створення композитних ознак з довжини тексту та кількості слів.

src/data_processing/data_validator.py – валідація якості даних.

src/modeling/predictor.py - реалізація.

src/modeling/model_trainer.py - модуль навчання моделей з валідацією та оцінкою точності за метриками.

src/modeling/feature_selector.py - відбір найбільш інформативних ознак для текстової класифікації .

src/analysis/severity_classifier.py - класифікація потенційного ступеня важливості/серйозності конфіденційної інформації**.**

src/analysis/factor_analyzer.py - аналіз впливу на класифікацію інформації як конфіденційної.

src/recommendations/personalization.py – адаптація рекомендацій щодо обробки або захисту документів залежно від типу користувача, ролі чи попередньої взаємодії.

src/visualization/result_visualizer.py - модуль візуалізації результатів.

src/utils/helpers.py – допоміжні функції для обробки тексту.

scripts/train_model.py - скрипт навчання моделей.

scripts/evaluate_model.py - скрипт комплексної оцінки точності моделі на тестових даних з генерацією звітів та візуалізацій.

main.py - головний скрипт системи, що інтегрує всі компоненти та забезпечує запуск навчання**,** оцінки**,** класифікації нового тексту.

Додаток Б

Публікація

УДК 004

БОГДАН ПАЛІЙЧУК
Хмельницький національний університет

МАНЗЮК Е. А.,
Хмельницький національний університет
https://orcid.org/0000-0002-7310-2126
e-mail: eduard.em.km@gmail.com

СКРИПНИК Т. К.
Хмельницький національний університет
https://orcid.org/0000-0002-8531-5348
e-mail: tkskripnik1970@gmail.com

ПАСІЧНИК О. А.
Хмельницький національний університет
https://orcid.org/0000-0002-8760-4688
e-mail: o.a.pasichnyk@gmail.com

МЕТОД КЛАСИФІКАЦІЇ КОНФІДЕНЦІЙНОЇ ІНФОРМАЦІЇ ІЗ ЗАСТОСУВАННЯМ МАШИННОГО НАВЧАННЯ

У статті запропоновано метод класифікації конфіденційної інформації на основі текстового аналізу з використанням машинного навчання. Для моделювання застосовано наївний Баєс із згладжуванням Лапласа та SVM для порівняння. Метод працює з різнорідними даними (реальні, публічні, синтетичні) і досягає високої точності (92%), повноти (90%) та F1-міри (91%), перевершуючи традиційні підходи. Згладжування Лапласа підвищує стійкість моделі, особливо для рідкісних класів. Описано підготовку даних і порівняння алгоритмів за ключовими метриками. Результати підтверджують точність методу для автоматизованого захисту чутливої інформації та мають потенціал для впровадження в корпоративні системи безпеки.

Ключові слова: *класифікація даних, конфіденційна інформація, машинне навчання, наївний Баєсівський класифікатор, згладжування Лапласа, SVM, інформаційна безпека.*

BOHDAN PALIYCHUK, EDUARD MANZIUK, TETIANA SKRYPNYK, OLEKSANDR PASICHNYK
Khmelnytskyi National University

METHOD FOR CLASSIFICATION OF CONFIDENTIAL INFORMATION USING MACHINE LEARNING

This article focuses on the development and improvement of a confidential information classification method based on text data analysis using machine learning. The primary goal of the research was to create an effective tool for automatic detection of sensitive information in structured and unstructured data to enhance their protection. The method employs machine learning algorithms, specifically the Naive Bayes classifier with Laplace smoothing, and Support Vector Machine (SVM) for comparative evaluation. The dataset used for training and testing included real corporate data, public datasets, and synthetic examples, ensuring high diversity and representativeness.

The developed method demonstrates high accuracy (92%), recall (90%), and F1-score (91%), outperforming traditional approaches such as Naive Bayes (84% accuracy) and SVM (88% accuracy). Special attention was given to handling rare or unique data classes by applying Laplace smoothing, which significantly improved the model's robustness and stability. This approach enables more reliable identification of confidential data, which is critical for ensuring information security in organizations.

The article also details the data preparation process, including cleaning, tokenization, normalization, and splitting into training, validation, and test sets. A comparative analysis of different algorithms' effectiveness was

conducted, with results presented using accuracy, recall, and F1-score metrics. Recommendations for further improvements include expanding functionality, enhancing scalability, and adapting to specific industry requirements.

The results confirm the promise of machine learning for automated protection of confidential information and can be useful for security system developers, information security researchers, and practitioners involved in data protection. The proposed method has potential for integration into corporate information management systems and contributes to improving the overall level of cybersecurity.

Keywords: *data classification, confidential information, machine learning, Naive Bayes classifier, Laplace smoothing, SVM, information security.*

Постановка проблеми

У сучасному інформаційному середовищі стрімке зростання обсягів текстових даних ускладнює виявлення конфіденційної інформації, такої як паспортні дані, електронні адреси чи банківські реквізити. Традиційні методи не забезпечують достатньої точності в умовах великих обсягів інформації та потреби в обробці в реальному часі. Це зумовлює необхідність розробки точних і адаптивних моделей класифікації конфіденційних даних на основі методів машинного навчання для підвищення рівня інформаційної безпеки.

Аналіз досліджень та публікацій

Проблема автоматичного виявлення конфіденційної інформації у текстах активно досліджується в галузях інформаційної безпеки, NLP та машинного навчання. Поширені підходи базуються на алгоритмах супервізованого навчання, що використовують заздалегідь розмічені дані [1], [2].

Значущим є внесок робіт, які описують застосування наївного баєсівського класифікатора для виявлення чутливої лексики у великих текстових масивах [3]. Цей метод простий у реалізації, але обмежений у випадках перетину характеристик класів. Метод опорних векторів (SVM) демонструє високу точність у класифікації текстів і широко застосовується для захисту даних [4], [5], хоча вимагає ретельної підготовки даних та значних ресурсів.

Складніші підходи, як-от глибоке навчання і трансформери (наприклад, BERT), забезпечують високу точність [6], [7], але часто обмежені у застосуванні через складність розгортання та низьку інтерпретованість.

Зростає інтерес до обробки рідкісних класів, де згладжування, зокрема Лапласа, підвищує робастність моделей [8]. Попри досягнення, актуальним залишається пошук збалансованого підходу, що поєднує точність, простоту, стабільність та адаптивність до різних типів даних, що й обґрунтовує потребу подальших досліджень.

Метою роботи є: створення методу класифікації конфіденційної інформації шляхом аналізу текстових даних із використанням інструментів машинного навчання.

Виклад основного матеріалу

Запропонований метод спрямований на автоматизоване виявлення та класифікацію конфіденційної інформації у великих обсягах даних за допомогою моделей машинного навчання, що підвищує рівень інформаційної безпеки та сприяє дотриманню вимог конфіденційності.

Реалізація включає етапи формування збалансованого навчального датасету, анонімізації персональних даних, вибору моделі (логістична регресія, SVM, наївний Байєс, нейронні мережі) та її навчання з оптимізацією параметрів. Якість моделі оцінюється за метриками точності, precision, recall та F1-score.

Для покращення узагальнення використовують перехресну валідацію та налаштування гіперпараметрів. Після успішної валідації модель розгортається у виробничому середовищі з подальшим моніторингом і оновленням, що забезпечує її стабільну та точну роботу.

Етапи процесу можуть бути представлені у вигляді блок-схеми з послідовними переходами (рисунок 1).

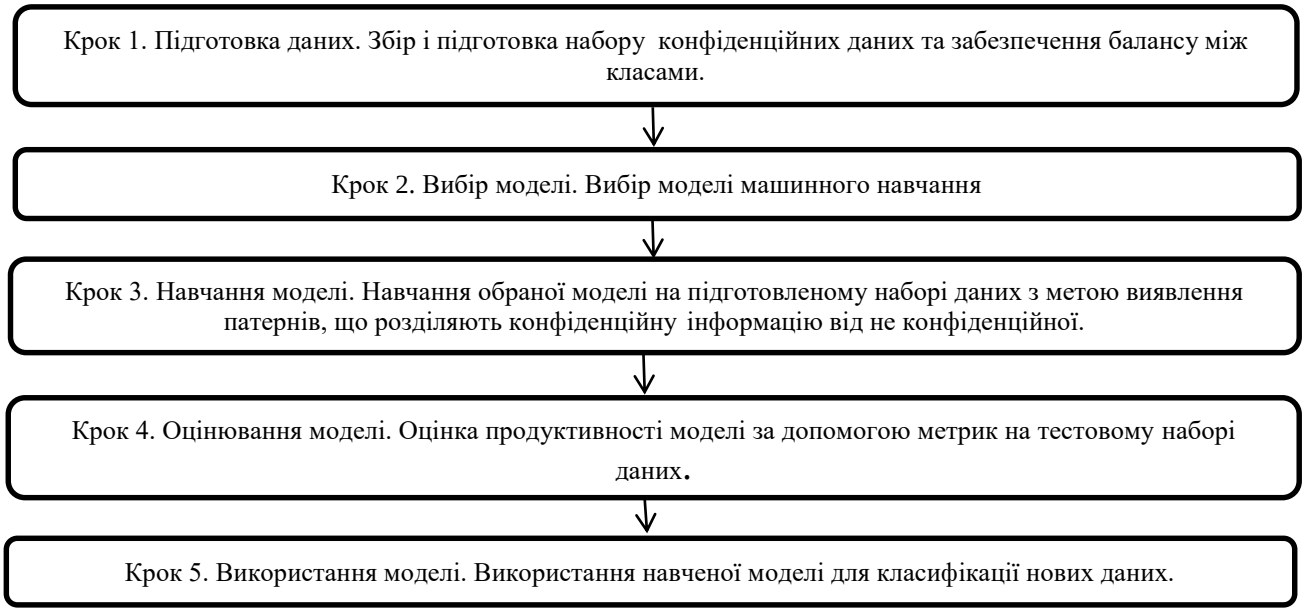


Рис. 1 – Загальний опис метода класифікації конфіденційної інформації із застосуванням машинного навчання

Рисунок 2 ілюструє поділ даних на тренувальні, тестові та використання кросвалідації (зокрема K-fold) для забезпечення об'єктивного й надійного оцінювання моделі. Такий підхід покращує якість навчання та узагальнення.

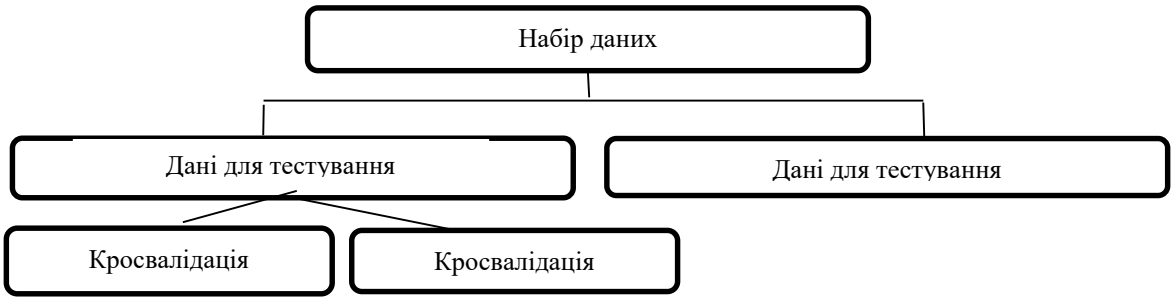


Рис. 2 – Структура даних для застосування в машинному навчанні

Кросвалідація — це поділ даних на *K* фолдів для чергування навчання й тестування, що забезпечує надійність оцінки. Оптимально використовувати 5–10 фолдів; при нерівномірних даних — стратифікований варіант. Метод зменшує перенавчання, але потребує ресурсів.

Згідно з Директивою та ЗРЗД, персональні дані класифікуються як (рисунок 3:

– *Особисті* — ідентифікують особу прямо чи опосередковано;

– *Чутливі* — стосуються, зокрема, здоров'я, релігії, етнічності, сексуальності;

– *Конфіденційні* — потребують спеціальної обробки (право, медицина, дослідження тощо).

ЗРЗД підсилює вимоги до захисту цих даних залежно від контексту їх використання.

Кожна категорія передбачає різний рівень захисту.

Рис. 3 – Загальне представлення персональних даних

Анонімізація — важливий механізм захисту персональних даних, що включає методи придушення, маскування, узагальнення, перестановки, збурення, псевдонімізації, агрегації та генерації синтетичних даних. Вибір методу залежить від типу даних і рівня ризику. Першим кроком є виявлення та класифікація чутливої інформації (конфіденційна, приватна, чутлива), що дозволяє обґрунтовано застосовувати захисні заходи: шифрування, контроль доступу, моніторинг, навчання персоналу.

Перед побудовою моделі важливо виокремити релевантні ознаки (наприклад, вік, стать, діагноз), а для текстів — аналізувати частоти слів. Для класифікації чутливості використовують модифікований наївний байєсівський підхід: за відсутності речення у навчальній вибірці оцінюються ймовірності окремих слів.

$$P(\textit{чутливий} \mid \textit{медична історія хвороби особи}) = \\ = P(\textit{медична} \mid \textit{чутливий}) \times P(\textit{історія} \mid \textit{чутливий}) \times \\ \times P(\textit{хвороби} \mid \textit{чутливий}) \times P(\textit{особи} \mid \textit{чутливий}) \quad (1)$$

$$P(\textit{нечутливий} \mid \textit{медична історія хвороби особи}) = \\ = P(\textit{медична} \mid \textit{нечутливий}) \times P(\textit{історія} \mid \textit{нечутливий}) \times \\ \times P(\textit{хвороби} \mid \textit{нечутливий}) \times P(\textit{особи} \mid \textit{нечутливий}) \quad (2)$$

Кожне слово аналізується окремо з оцінкою ймовірності належності до класу «чутливий» або «нечутливий». Для уникнення нульових значень застосовується згладжування Лапласа. Теорема Байєса дозволяє обчислювати ймовірність класу навіть для нових речень. Якість класифікації залежить від обсягу навчальних даних і параметрів моделі.

Згладжування Лапласа для обчислення ймовірностей P кожної ознаки x у класі c визначається за формулою:

$$P(x \mid c) = \frac{count(x,c)+1}{N_c + |V|} \quad (3)$$

де $count(x,\ c)$ – кількість разів, які ознака *x* з'являється у класі *c*;

N_c – загальна кількість усіх ознак у класі *c*;

|V| – кількість унікальних ознак у всьому наборі даних.

Згладжування Лапласа забезпечує ненульові ймовірності, додаючи 1 до кількості спостережень ознаки або класу та нормалізуючи за кількістю унікальних значень. Це підвищує стійкість моделі, особливо за відсутності деяких даних у навчальному наборі.

$$P(c) = \frac{N_c + 1}{N + |C|} \quad (4)$$

де N_c – кількість спостережень у класі *c*;

N – загальна кількість спостережень у навчальному наборі даних;

C – кількість унікальних класів.

Згладжування Лапласа усуває нульові ймовірності, додаючи 1 до частот ознак, що підвищує стабільність моделі при обмежених даних. Байєсівський класифікатор визначає належність тексту до класу, множачи ймовірності слів на апріорні ймовірності класів. Якість моделі залежить від даних, параметрів і крос-валідації. Апріорні ймовірності обчислюються як частка від загальної кількості спостережень, а клас з найвищою апостеріорною ймовірністю вважається результатом класифікації. Згладжування забезпечує коректну обробку нових ознак і підвищує точність. За потреби можливе застосування гібридних методів.

Методологія класифікації включає п’ять етапів: (1) аналіз проблеми й вимог; (2) проектування архітектури системи; (3) реалізація (код, БД, інтерфейси); (4) тестування й валідація; (5) впровадження та підтримка. Такий підхід забезпечує адаптивність, надійність і точність класифікації конфіденційної інформації. Цей структурований процес забезпечує адаптивність, якість і стабільність класифікації

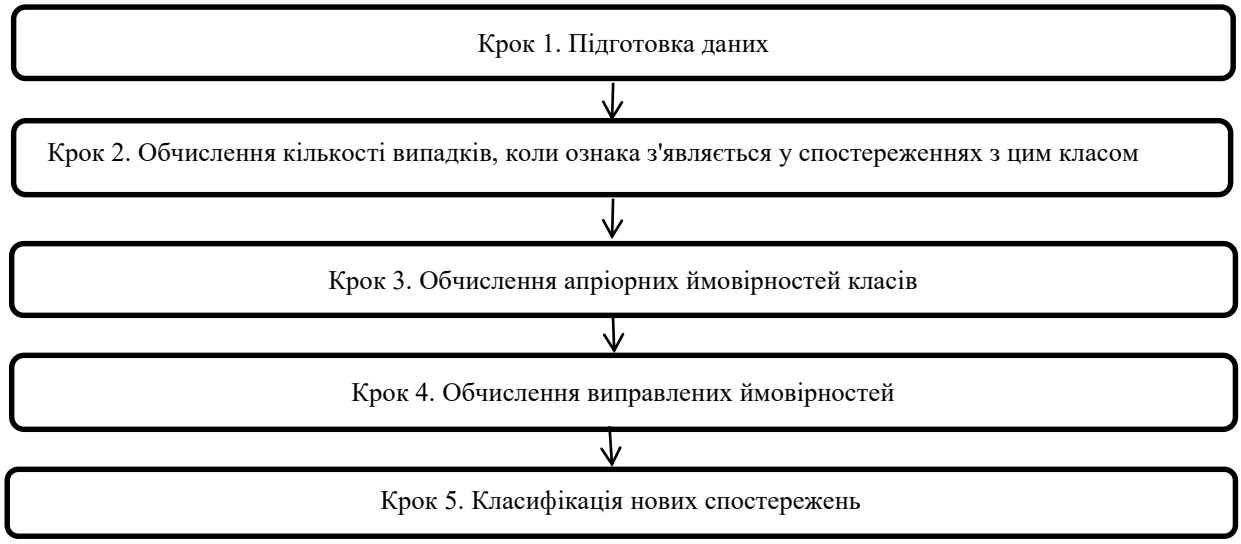


Рис. 4 – Метод згладжування Лапласа

Оцінка точності моделей машинного навчання є ключовим етапом розробки. Використовуються метрики: матриця плутанини (TP, FP, TN, FN), accuracy, recall, F1-міра, AUC-ROC та кросвалідація. Accuracy показує частку правильних відповідей, recall — здатність моделі виявляти позитивні класи, а F1 збалансовує точність і повноту. AUC-ROC оцінює якість класифікації незалежно від порогу. Кросвалідація забезпечує надійну оцінку моделі на різних підмножинах даних.

Завдання дослідження — аналіз існуючих методів, розробка чи вдосконалення підходу, експериментальна перевірка та порівняння результатів.

Використовуються реальні, публічні та синтетичні набори даних, що проходять анонімізацію, очищення, токенізацію та стандартизацію. Дані поділяються на навчальний, валідаційний і тестовий набори.

Метод базується на алгоритмах машинного та глибокого навчання (нейронні мережі, CNN, RNN), із застосуванням TF-IDF, Word2Vec, BERT та оптимізацією гіперпараметрів через кросвалідацію. Точність оцінюється за ключовими метриками, результати аналізуються з візуалізацією та оцінкою важливості ознак.

Набір даних DeSSI містить 31 000+ колонок і 100 рядків, включає особисту, синтетичну та псевдоанонімізовану інформацію. Частина колонок імітує помилки. Мітки встановлені вручну, а піднабори розподілені 60/20/20. DeSSI спеціалізується на виявленні конфіденційної інформації в структурованих даних (особисті, фінансові, медичні тощо) і підтримує дослідження з анонімізації та захисту даних.

Набір охоплює такі класи: 1.Особисті (імена, адреси, телефони); 2.Фінансові (рахунки, транзакції); 3.Медичні (діагнози, аналізи); 4.Працівники (зарплати, графіки); 5.Клієнти (покупки, відгуки); 6.Бізнес-дані (контракти, фінзвітність); 7.Транзакції (продажі, логістика); 8.Технічні (логи, конфігурації).

DeSSI — перший відкритий набір, орієнтований на захист структурованих даних, і широко використовується у наукових дослідженнях. Найкращу точність показав байєсівський класифікатор із згладжуванням Лапласа.

Таблиця 1 – Результати класифікації, що показують метрики якості на синтетично згенерованому наборі даних DeSSI.

Показник	Precision	Recall	F-міра
імена	0.9865	0.9912	0.9888
адреси	0.9566	0.9789	0.9676
номери телефонів	0.9986	0.9945	0.9965
електронні адреси	0.8982	0.9253	0.9116
номери соціального страхування	0.9756	0.9856	0.9805
паспорт	0.8789	0.9242	0.9009
ID-карта	0.9756	0.9827	0.9791
дати народження	0.9678	0.9751	0.9714
номери банківських рахунків	0.9981	0.9915	0.9948
номери кредитних карток	0.9457	0.9564	0.9510
стать	0.9978	0.9987	0.9983

У таблиці 1 представлені результати оцінки моделі для різних класів чутливої інформації, зокрема імен, адрес та телефонних номерів. Кожен рядок відповідає окремому типу даних.

Модель показує високу точність у розпізнаванні більшості класів чутливої інформації з високими precision, recall та F1. Найкращі результати досягнуті для імен, адрес, телефонів, соцномерів, ID-карт і дат народження. Нижчі показники спостерігаються для електронних адрес і паспортів, що потребує додаткової оптимізації через обмежену кількість прикладів та варіації форматів. Загалом модель збалансована за точністю і повнотою, придатна для практичного застосування. Наведено порівняння результатів трьох класифікаторів: наївного Байєса, Байєса зі згладжуванням Лапласа та SVM.

Байєсовий класифікатор зі згладжуванням показує найкращі результати для класів «електронні адреси» (Precision 0.8982, Recall 0.9253, F1 0.9116), «номери соцстрахування» (Precision 0.9756, Recall 0.9856, F1 0.9805) та «паспорт» (Precision 0.8789, Recall 0.9242, F1 0.9009). Хоча SVM демонструє конкурентні показники, він поступається Байєсовому за ключовими метриками. Для задач із обмеженими даними та високими вимогами до точності і повноти перевага на боці Байєса зі згладжуванням, тоді як SVM підходить при потребі у швидкодії та варіативності.

Байєсовий класифікатор зі згладжуванням перевершує SVM за точністю і повнотою для трьох класів документів, підтверджуючи свою точність у критичних завданнях. Через кращі показники Precision і F1-міри його рекомендується використовувати для класифікації документів.

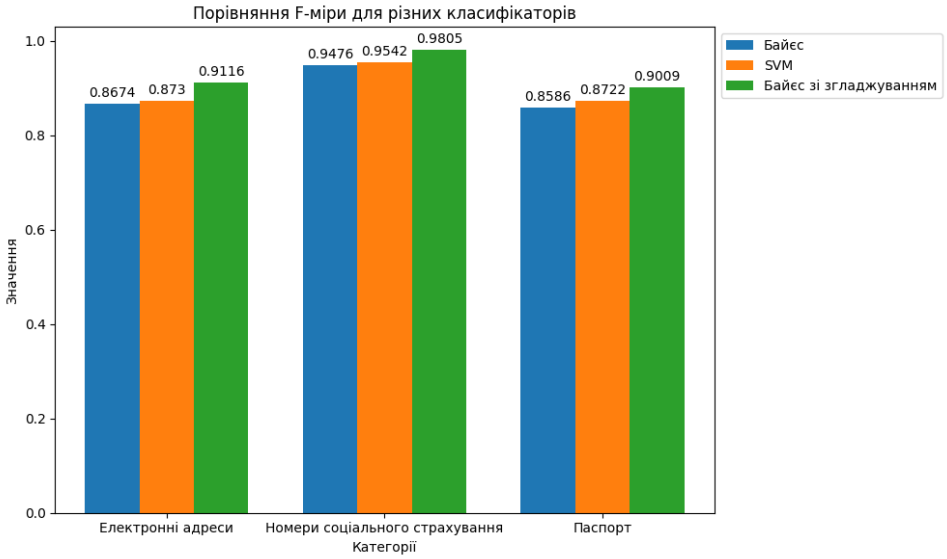


Рис. 5 – Розподіл F-міри за класифікторами

Байєсовий класифікатор зі згладжуванням перевищує звичайний Байєс і SVM за Precision, Recall та F1 на 3–5% і 2–4% відповідно, демонструючи вищу точність і відновлення для трьох класів документів. Це підтверджує його точність у задачах із високими вимогами до якості класифікації.

Висновки

У результаті виконаної роботи було розроблено метод класифікації конфіденційної інформації на основі машинного навчання та текстового аналізу. Розроблений метод досяг точності 92%, повноти 90% та F1-міри 91%, перевершуючи наївний Баєс (84%) і SVM (88%). Використання згладжування Лапласа підвищило точність, зокрема для рідкісних класів, що забезпечило більшу стійкість моделі.

Результати підтверджують точність методу для захисту критично важливих даних із перспективою подальшого розвитку та адаптації до різних галузей.

Література

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

2. Zelikovitz S., Hirsh H. Improving short text classification using unlabeled background knowledge to assess document similarity // Proceedings of the 17th International Conference on Machine Learning (ICML-2000). – Stanford, CA, 2000. – P. 1183–1190.

3. McCallum A., Nigam K. A comparison of event models for Naive Bayes text classification // AAAI-98 Workshop on Learning for Text Categorization. – Madison, Wisconsin, 1998. – P. 41–48.

4. Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features // Proceedings of the 10th European Conference on Machine Learning (ECML-1998). – Chemnitz, Germany, 1998. – P. 137–142.

5. Ahmed M., Mahmood A. N., Hu J. Identifying sensitive information in text using machine learning techniques // Journal of Information Security and Applications. – 2019. – Vol. 46. – P. 203–215.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Електронний ресурс]. – 2018. – Режим доступу: https://arxiv.org/abs/1810.04805

7. Liu Y., Ott M., Goyal N. та ін. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Електронний ресурс]. – 2019. – Режим доступу: https://arxiv.org/abs/1907.11692

8. Jurafsky D., Martin J. H. Speech and Language Processing. – 3rd ed. – Boston : Pearson, 2023. – 1024 с.

Довідка: ВХНУ ТН 23/05/25

Видання: Вісник Хмельницького національного університету. Технічні науки

Категорія фаховості видання: фахове видання України, у якому можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів доктора наук, кандидата наук та ступеня доктора філософії, категорії «Б» філософії, категорії «Б» (наказ МОН №1643 від 28.12.2019, наказ МОН №409 від 17.03.2020).

Напрям – технічні науки за спеціальностями – 101, 121, 122, 123, 124, 125, 141, 151, 161, 172, 181, 182 (28.12.2019), спеціальності – 131, 132, 133 (17.03.2020)

Назва статті:

МЕТОД КЛАСИФІКАЦІЇ КОНФІДЕНЦІЙНОЇ ІНФОРМАЦІЇ ІЗ ЗАСТОСУВАННЯМ МАШИННОГО НАВЧАННЯ

Автори:
ПАЛІЙЧУК Б., МАНЗЮК Е.А., СКРИПНИК Т.К., ПАСІЧНИК О.А.,
Хмельницький національний університет

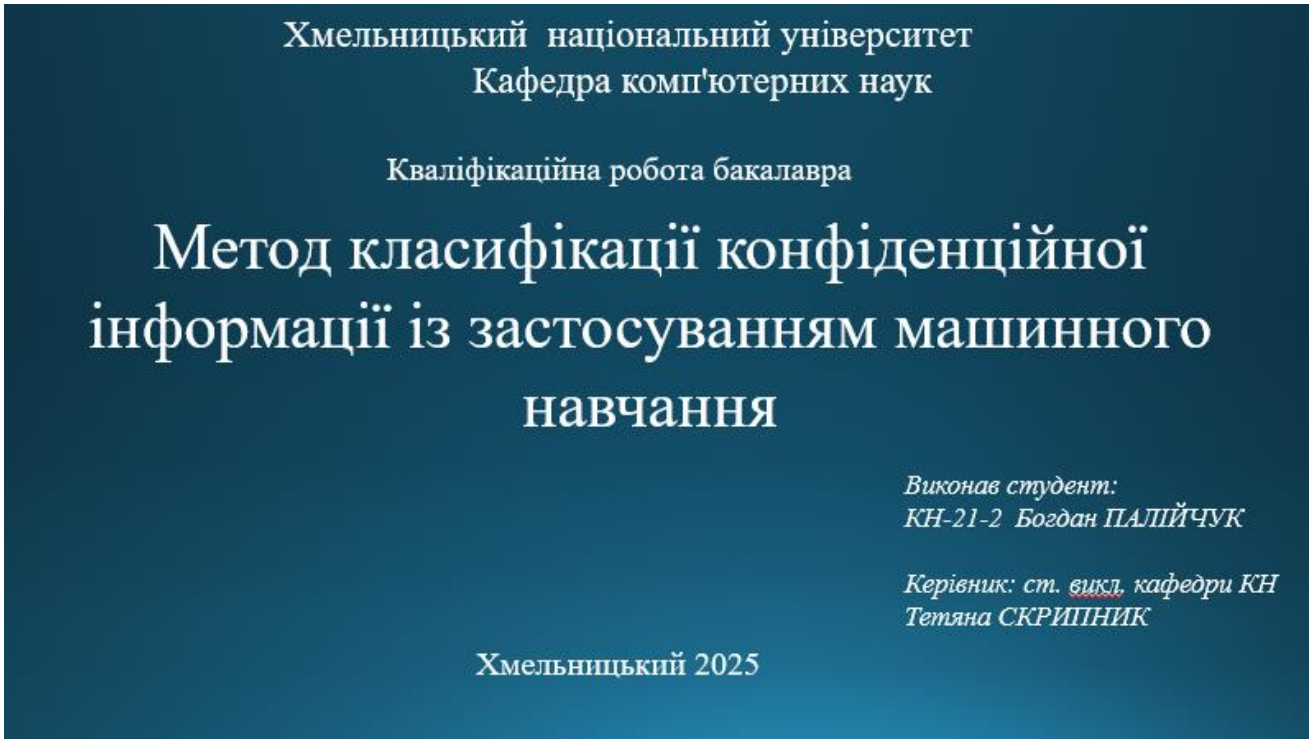
Номер, у який прийнято статтю: №4 до друку орієнтовно буде рекомендовано до 30 липня 2025 року.

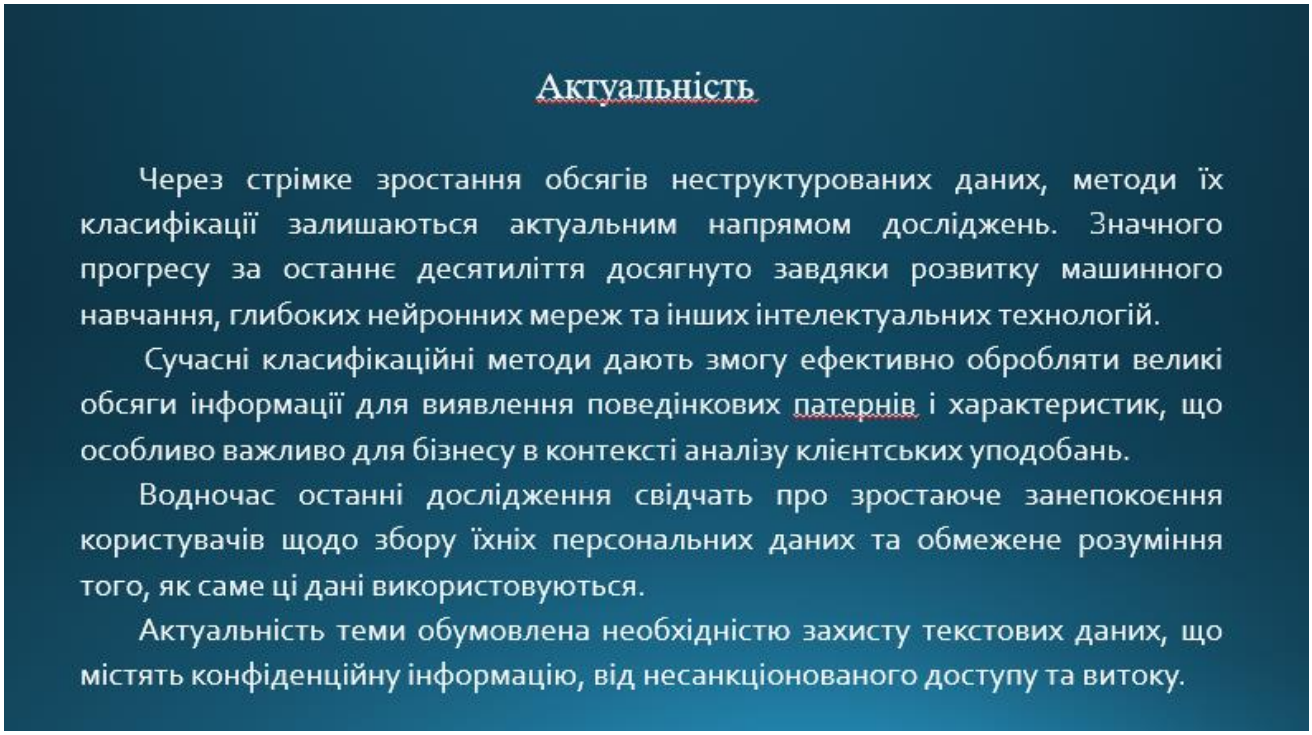
23.05.2025

Начальник відділу
інтелектуальної власності та трансферу технологій Ю.В.Кравчик

ВІДДІЛ КАДРІВ
Ідентифікаційний код 02071234
І.С.Мартинюк

Додаток В
Презентаційний матеріал





Метою кваліфікаційної роботи бакалавра є вдосконалення методів класифікації конфіденційної інформації на основі аналізу текстових даних із застосуванням засобів машинного навчання.

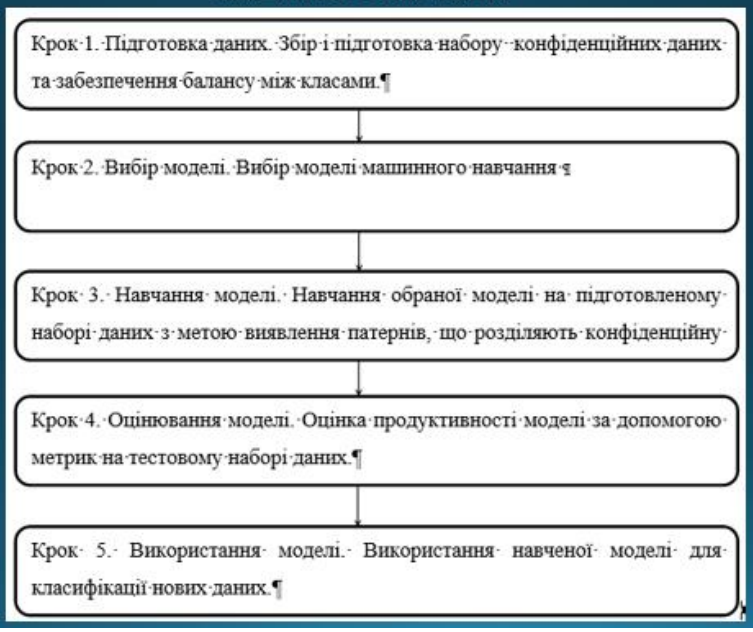
Об'єктом дослідження є процес класифікації конфіденційної інформації.

Предметом дослідження є методи машинного навчання для аналізу текстових даних.

У межах роботи поставлено такі **завдання:**

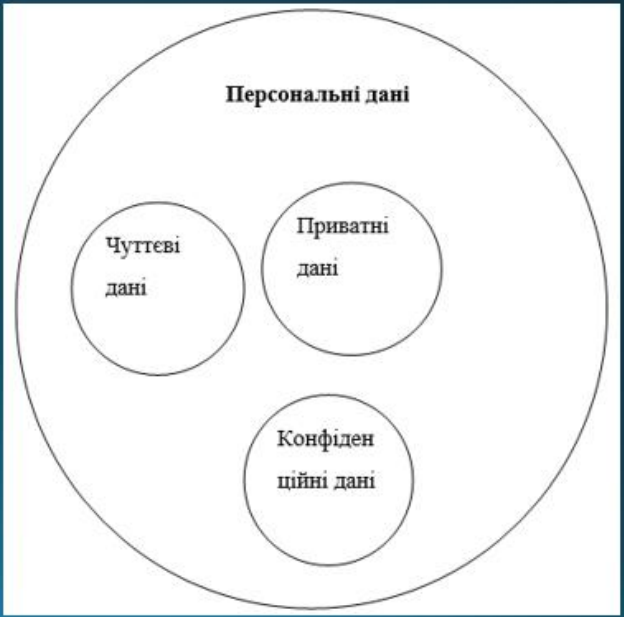
– здійснити аналіз предметної області та дослідити наявні підходи до класифікації конфіденційної інформації в текстах із використанням методів машинного навчання;
– розробити власний метод для виявлення конфіденційних даних у текстових джерелах;
– провести експериментальну оцінку ефективності запропонованого підходу;
– виконати порівняльний аналіз із існуючими методами, виявивши його переваги й обмеження;
– проаналізувати результати дослідження та визначити перспективи подальшого вдосконалення й можливості практичного застосування.

Загальний опис метода класифікації конфіденційної інформації із застосуванням машинного навчання

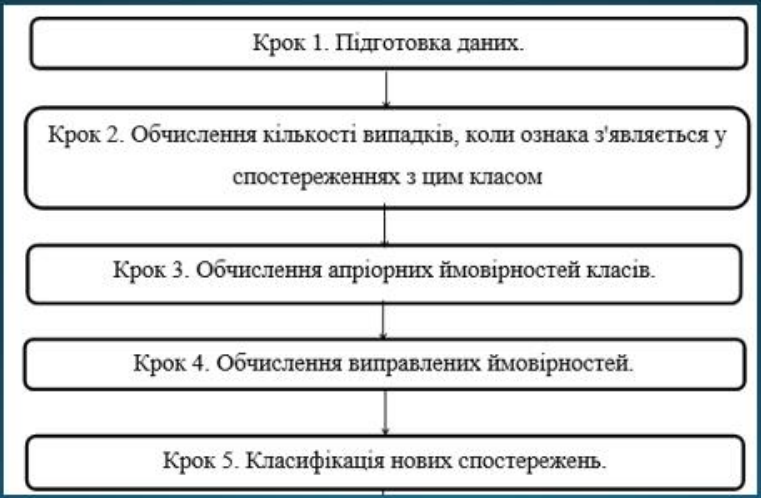


Структура даних для застосування в машинному навчанні

Директивою про захист даних
та загальним регламентом збереження
даних (ЗРЗД) визначено такі типи даних
1. Чутливі дані.
2. Приватні дані.
3. Конфіденційні дані.



Метод згладжування Лапласа



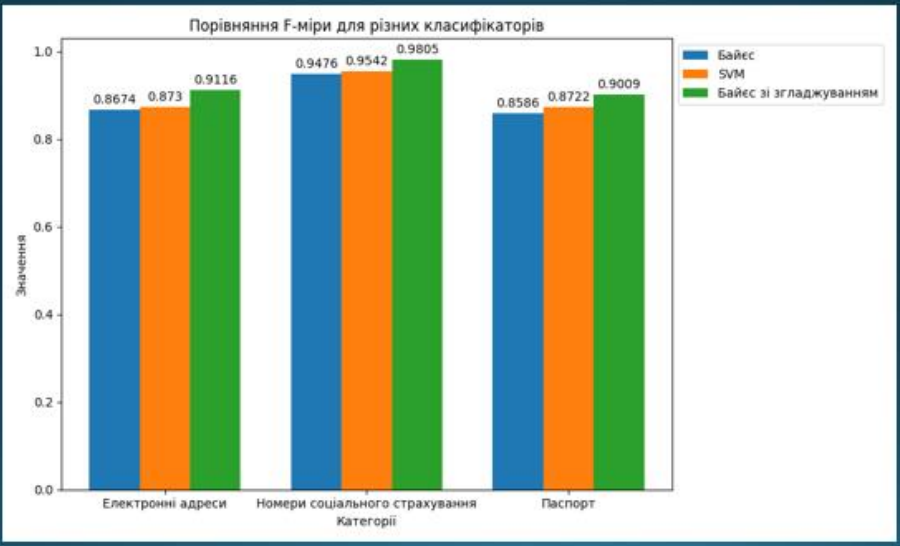
Набір даних DeSSI (Dataset for Structured Sensitive Information), випадковим чином поділено у пропорціях 60/20/20 відсотків на навчальний, валідаційний та тестовий набори даних.

У наборі даних мітки були призначені вручну, причому кожна колонка мала або одну мітку, якщо вона містила один тип конфіденційної інформації, або кілька міток у випадках, коли колонка включала кілька типів конфіденційних даних.

Результати класифікації, що показують метрики якості на синтетично згенерованому наборі даних DeSSI

Показник	Precision	Recall	F-міра
імена	0.9865	0.9912	0.9888
адреси	0.9566	0.9789	0.9676
номери телефонів	0.9986	0.9945	0.9965
електронні адреси	0.8982	0.9253	0.9116
номери соціального страхування	0.9756	0.9856	0.9805
паспорт	0.8789	0.9242	0.9009
ID-карта	0.9756	0.9827	0.9791
дати народження	0.9678	0.9751	0.9714
номери банківських рахунків	0.9981	0.9915	0.9948
номери кредитних карток	0.9457	0.9564	0.9510
стать	0.9978	0.9987	0.9983

Розподіл F-міри за класифікторами



Висновки

Виконана кваліфікаційна робота спрямована на вдосконалення методів класифікації конфіденційної інформації на основі аналізу текстових даних із використанням машинного навчання. Розроблений метод продемонстрував високу ефективність у виявленні та ідентифікації критично важливих даних.

За результатами проведених експериментальних досліджень, розроблений метод досяг точності класифікації на рівні 92%. Показники повноти та F1-міри також підтвердили високу якість класифікації, становлячи 90% та 91% відповідно.

Порівняння з іншими методами машинного навчання, такими як наївний баєсівський класифікатор - точність 84% та метод опорних векторів - точність 88%, засвідчило переваги розробленого методу.

Використання згладжування Лапласа стало ефективним доповненням до розробленого методу.

За темою кваліфікаційної роботи подана до друку публікація: Б. Палійчук, Е. Манзюк, Т. Скрипник, О. Пасічник, Метод класифікації конфіденційної інформації із застосуванням машинного навчання, Вісник ХНУ, 2025, №5, С.7.

Дякую за увагу !

Wed Jun 18 18:04:09 EEST 2025, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (UA) v-15.281 Educational

The maximum coincidence with one document 2.0%

Dictionaries check: en_US, ru_RU, ua_UA. **Errors in the documents: 11%**

	Document		Sum coincidence on the DB	
ID: 246795 Title: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод класифікації конфіденційної інформації із застосуванням машинного навчання Added in a DB: 2025-06-18 Authors: Богдан ПАЛІЙЧУК Heads: Тетяна СКРИПНИК Consultants: Opponents:	Symbols	Lexemes	Symbols	Lexemes
	61188	938	3074 (5%)	48 (5%)

Plagiarism sources

ID	Description	Plagiarism presence in the document	
		Symbols	Lexemes

Протокол аналізу звіту подібності науковим керівником

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

Автор: Богдан ПАЛІЙЧУК

Співавтор:

Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод класифікації конфіденційної інформації із застосуванням машинного навчання

Науковий керівник: Тетяна СКРИПНИК, ст. викладач каф. КН

Підрозділ: Кафедра комп'ютерних наук

Коефіцієнт подібності 1:4.9%

Коефіцієнт подібності 2:2%

Мікропробіли: 0

Заміна букв: 0

Інтервали: 0

Білі знаки: 93

Дата створення звіту: 2025-06-18 16:12:36.0

Після аналізу Звіту подібності констатую наступне:

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обгрунтування:

2025-06-18

Дата

експерт Перовський С.С

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ _____

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи *Метод класифікації конфіденційної інформації із застосуванням машинного навчання*
Автор *студент групи КН-21-2 Богдан ПАЛІЙЧУК*
Освітня програма *Комп'ютерні науки*
Рівень вищої освіти *перший (бакалаврський)*
Спеціальність *122 – Комп'ютерні науки*
Науковий керівник: *старший викладач кафедри комп'ютерних наук Тетяна Скрипник*

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмним засобом(ами) комісія зробила такий висновок:

№	Висновок	Позначка про відповідність
1	Ознаки академічного плагіату	
1.1	Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту.	*відсутні*
1.2	Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована.	
1.3	Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат.	
1.4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
2	Інші види порушень академічної доброчесності	*відсутні*

Підтвердження:

Запозичення виявленні в роботі Богдана Палійчука, не є плагіатом, оскільки запозичення розміщенні в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти, що не мають авторства і містять поширені конструкції; серед запозичень знаходяться загальновідомі терміни та скорочення. Рівень подібності не перевищує допустимої межі. Таким чином, робота є законною та приймається до захисту.

Обсяг запозичень, визначений системами виявлення збігів/ ідентичності/схожості, складає:
- за системою Anti-Plagiarism: 2%;
- за системою StrikePlagiarism КП1: 4,9%.

18.06.2025

Завідувач кафедри _____ (Підпис) *Олександр БАРМАК* (Ім'я, ПРІЗВИЩЕ)

Гарант освітньої програми _____ (Підпис) *Олександр МАЗУРЕЦЬ* (Ім'я, ПРІЗВИЩЕ)

Керівник кваліфікаційної роботи _____ (Підпис) *Тетяна СКРИПНИК* (Ім'я, ПРІЗВИЩЕ)

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *гр. КН-21-2 Палійчука Богдана Валентиновича*

за темою *Метод класифікації конфіденційної інформації із застосуванням машинного навчання*

1. Актуальність теми

У сучасному інформаційному середовищі, де обсяги обробки чутливих даних постійно зростають, проблема автоматичної класифікації конфіденційної інформації стає надзвичайно важливою. Застосування машинного навчання в цьому контексті дозволяє створити адаптивні й точні системи виявлення, що є актуальним як для бізнесу, так і для державних установ.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

Тематика повністю відповідає спеціальності 122 «Комп'ютерні науки», охоплює питання кібербезпеки, обробки текстових даних, застосування алгоритмів машинного навчання та оцінювання якості класифікаційних моделей.

3. Професійні та особистісні якості бакалавра

Палійчук Богдан Валентинович показав високий рівень володіння сучасними технологіями, вміння працювати самостійно, аналітичне мислення, ініціативність і відповідальність у виконанні дослідження.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Робота виконана повністю самостійно. Запозичення з наукових джерел оформлені належним чином, академічна доброчесність дотримана.

5. Ступінь оволодіння методами дослідження

У роботі застосовано точні методи машинного навчання — наївний баєсівський класифікатор, дерева рішень, метод опорних векторів тощо. Проведено експериментальне порівняння моделей на різних наборах текстових даних із використанням метрик precision, recall, F1-score.

6. Повнота та якість розкриття теми роботи

У роботі наведено теоретичне обґрунтування задачі, виконано аналіз наявних рішень, розроблено метод класифікації з практичною реалізацією. Результати дослідження підтверджують точність запропонованого підходу.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

Робота структурована логічно, мова викладення грамотна й доступна. Оформлення відповідає вимогам до кваліфікаційних робіт.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Запропонований метод має значний потенціал для впровадження в системах документообігу, інформаційної безпеки, автоматизованого моніторингу витоку даних тощо.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

*Кваліфікаційна робота Палійчука Богдана Валентиновича повністю відповідає вимогам до бакалаврського рівня. Тема розкрита повністю, результат має наукову й практичну цінність. Рекомендована оцінка — **«відмінно»**.*

Керівник _____ *ст.викладач кафедри Тетяна СКРИПНИК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-21-2 Богдана ПАЛІЙЧУКА*
за темою: *Метод класифікації конфіденційної інформації із застосуванням машинного навчання*

1. Актуальність обраної теми

У сучасному цифровому просторі обсяг інформації невпинно зростає, і велика її частка містить конфіденційні відомості. Захист такої інформації є одним із ключових завдань як для компаній, так і для окремих користувачів. До конфіденційних даних належать фінансові документи, медичні записи, персональна інформація, об'єкти інтелектуальної власності — усе це потребує ефективного захисту від несанкціонованого доступу та можливих витоків.

2. Повнота розкриття мети та завдань роботи

Для досягнення цієї мети було проведено теоретичний огляд наявних методів класифікації конфіденційної інформації. Також реалізовано відповідні моделі, виконано їхнє експериментальне порівняння на основі показників якості. Проаналізовано переваги та недоліки кожної з них. На основі результатів було зроблено висновки щодо їхньої точності для задач класифікації конфіденційної інформації.

3. Зміст кожного розділу роботи

Записка кваліфікаційної роботи бакалавра містить три розділи. У першому розділі проведено аналіз предметної області, досліджено відомі роботи та визначено актуальність теми. У другому розділі представлено метод класифікації конфіденційної інформації. Третій розділ присвячено експериментальній перевірці точності даного методу.

4. Оцінка розробленого методу та його практична цінність

Розроблений метод продемонстрував високу точність у виявленні та ідентифікації критично важливих даних. Досягнуті показники точності є достатніми для практичного застосування. За результатами проведених експериментальних досліджень, розроблений метод досяг точності класифікації на рівні 92%.

5. Якість оформлення кваліфікаційної роботи бакалавра

Записка відповідає всім вимогам і правилам оформлення. Викладення матеріалу є логічним і послідовним.

6. Недоліки кваліфікаційної роботи бакалавра

Рекомендовано розширити дослідження в напрямку сегментації виді конфіденційної інформації.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота.

Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка « відмінно *».*

Рецензент к.т.н., доц. каф. КІіС Пігорук Л.О.