https://doi.org/10.31891/2307-5732-2025-357-13
УДК 004.4

БОГДАН ПАЛІЙЧУК
Хмельницький національний університет
e-mail: tkskripnik1970@gmail.com
МАНЗЮК ЕДУАРД
Хмельницький національний університет
https://orcid.org/0000-0002-7310-2126
e-mail: eduard.em.km@gmail.com
СКРИПНИК ТЕТЯНА
Хмельницький національний університет
https://orcid.org/0000-0002-8531-5348
e-mail: tkskripnik1970@gmail.com
ПАСІЧНИК ОЛЕКСАНД
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 ефективність 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.*

Стаття надійшла до редакції / Received 10.07.2025
Прийнята до друку / Accepted 15.08.2025

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 ефективності 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 ефективну роботу.

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

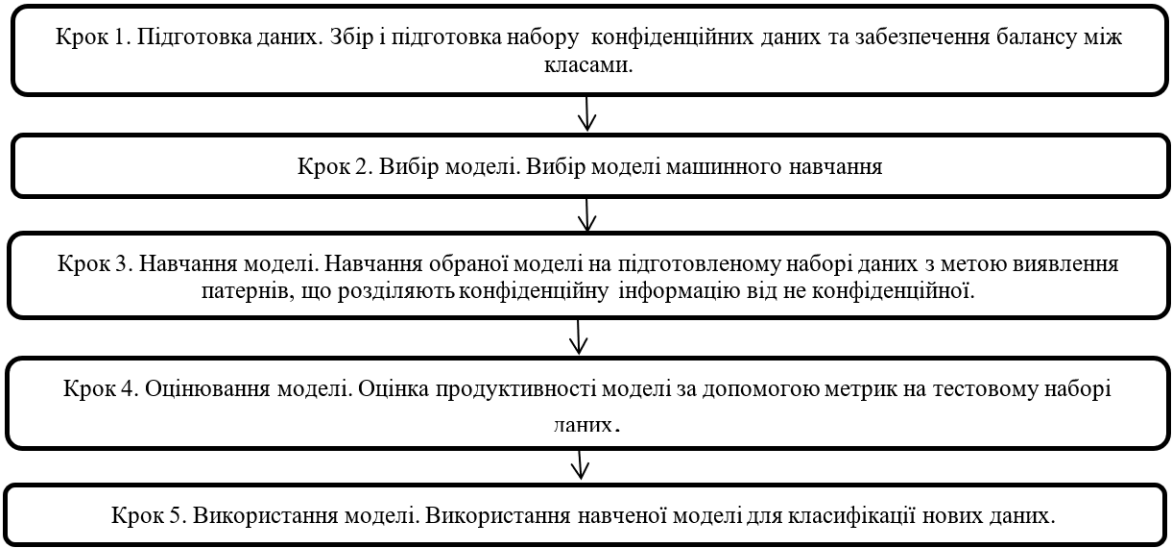

Рис.1. Загальний опис метода класифікації конфіденційної інформації із застосуванням машинного навчання

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.

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.

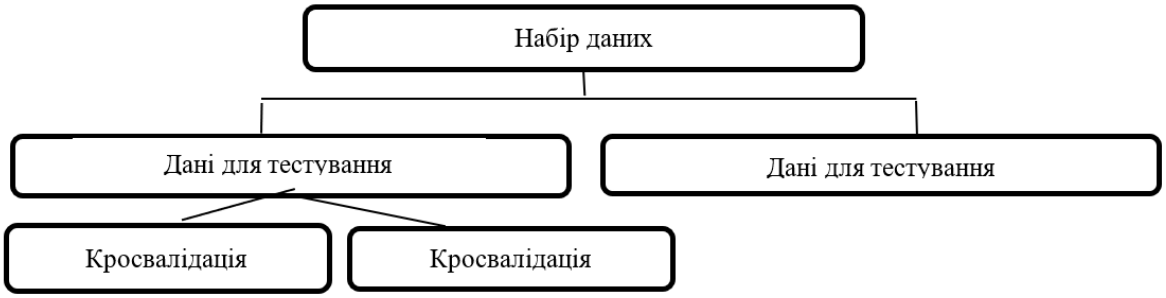

Рис.2. Структура даних для застосування в машинному навчанні

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.

$$\begin{aligned} &P(чутливий \mid медична\ історія\ хвороби\ особи) = \\ &= P(медична \mid чутливий) \times P(історія \mid чутливий) \times \\ &\times P(хвороби \mid чутливий) \times P(особи \mid чутливий) \end{aligned} \tag{1}$$

$$\begin{aligned} &P(нечутливий \mid медична\ історія\ хвороби\ особи) = \\ &= P(медична \mid нечутливий) \times P(історія \mid нечутливий) \times \\ &\times P(хвороби \mid нечутливий) \times P(особи \mid нечутливий) \end{aligned} \tag{2}$$

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:

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

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 ефективність класифікації конфіденційної інформації. Цей структурований процес забезпечує адаптивність, якість і стабільність класифікації

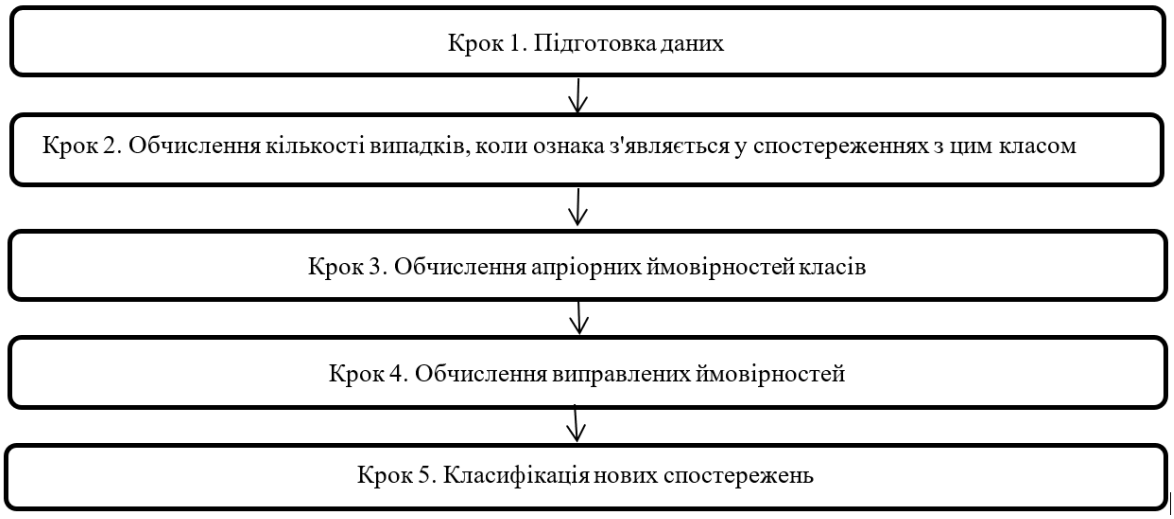

Рис.4. Метод згладжування Лапласа

Оцінка ефективності 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. Ефективність 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 ефективність 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 ефективність у критичних завданнях. Через кращі показники Precision і F1-міри його рекомендується використовувати для класифікації документів.

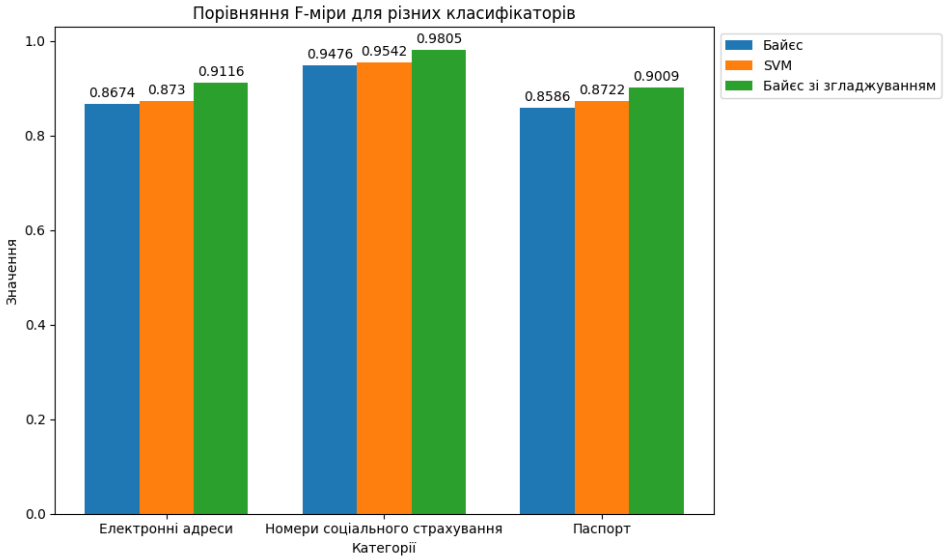

Рис.5. Розподіл F-міри за класифікторами

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 ефективність у задачах із високими вимогами до якості класифікації.

Висновки

У результаті виконаної роботи було розроблено ефективний 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 ефективність 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.