МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

ЧІП Святослав Андрійович

Метод автоматичної класифікації авторів текстів на основі машинного навчання / Method for automated text author classification based on machine learning

спеціальність: 122 - Комп'ютерні науки

освітньо-професійна програма - Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КНм-21

С.А. Чіп

_____

Науковий керівник:

к.т.н., доцент

Х.В. Ліп'яніна-Гончаренко

_____

Кваліфікаційну роботу

допущено до захисту:

«___» _____ 20___ р.

В.о. завідувача кафедри

_____ Н.В. Дзюбановська

ТЕРНОПІЛЬ - 2025

Факультет комп'ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

Освітній ступінь «магістр»

спеціальність: 122 – Комп'ютерні науки

освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ

В.о. завідувача кафедри

_____ Н.М. Васильків

«____» _____20__р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

Чіп Святослав Андрійович

(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи

Метод автоматичної класифікації авторів текстів на основі машинного навчання / Method for automated text author classification based on machine learning

керівник роботи к.т.н., доцент Х.В. Ліп'яніна- Гончаренко

затверджені наказом по університету від 20 грудня 2024 року № 938.

2. Строк подання студентом закінченої кваліфікаційної роботи 1 грудня 2025 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити

- огляд предметної області та аналіз задачі автоматичної класифікації авторів текстів у системах інтелектуального аналізу даних та комп'ютерної лінгвістики;
- аналіз стилометричних підходів до представлення тексту, зокрема словних і символьних n-грам, TF–IDF-векторизації та статистичних мовних ознак;
- огляд класичних методів машинного навчання для задач авторської атрибуції**,** включаючи наївний баєсівський класифікатор, логістичну регресію, метод опорних векторів та їх властивості;
- огляд імовірнісних підходів та методів калібрування ймовірностей у багатокласовій класифікації текстів;
- постановка задачі багатокласової автоматичної класифікації авторів текстів з урахуванням розрідженості ознак і дисбалансу класів;
- розробка методу автоматичної класифікації авторів текстів на основі регуляризованих лінійних моделей машинного навчання;
- дослідження та застосування імовірнісного ансамблювання моделей (зокрема поєднання Multinomial Naive Bayes і логістичної регресії) для підвищення якості прогнозування;
- реалізація програмної моделі, моделювання процесу навчання та проведення експериментальних досліджень на відкритих текстових корпусах;

5. Перелік графічного матеріалу у роботі

- схема методу автоматичної класифікації авторів текстів на основі машинного навчання;
- графіки порівняльного аналізу ефективності базових моделей та імовірнісного ансамблю за основними метриками якості класифікації.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 20 грудня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2025 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2025 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 20.05.2025 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 28.10. 2025 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 11.11.2025 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 25.11.2025 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту.	до 1.12.2025 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 04.12.2025 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.12. 2025 р.	

Студент _____ С.А. Чіп
підпис

Керівник роботи _____ к.т.н., доцент Х.В. Ліп’яніна- Гончаренко

РЕЗЮМЕ

Кваліфікаційна робота на тему «Метод автоматичної класифікації авторів текстів на основі машинного навчання» на здобуття освітнього ступеня «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньо-професійної програми «Комп'ютерні науки» виконана на 56 сторінках і містить 8 ілюстрацій, таблиці, додатки та 32 список використаних джерел.

Метою роботи є розроблення та експериментальне обґрунтування методу автоматичної атрибуції авторства текстів на основі машинного навчання з використанням стилометричних ознак, регуляризованих лінійних моделей і каліброваних імовірнісних ансамблів для забезпечення інтерпретованості та стабільної узагальнювальної здатності.

У роботі виконано огляд сучасних підходів до авторської атрибуції та сформовано вимоги до подання текстових даних. Запропоновано метод автоматичної класифікації авторів на основі TF–IDF-подань словних і символьних n-грам у поєднанні з Multinomial Naive Bayes і багатокласовою логістичною регресією з регуляризацією. Проведено порівняльне експериментальне дослідження базових моделей та імовірнісного ансамблю NB+LR з підбором ваг за мінімумом валідаційної крос-ентропії.

Отримані результати підтверджують ефективність запропонованого методу та доцільність застосування ансамблювання для підвищення узгодженості імовірнісних прогнозів без зниження точності класифікації.

Практичне значення роботи полягає у можливості використання методу в системах перевірки академічної доброчесності, цифрової гуманітаристики, медіа-форензіки, інформаційної безпеки та інших системах аналізу текстів.

Ключові слова: АВТОРСЬКА АТРИБУЦІЯ, АНАЛІЗ ТЕКСТІВ, МАШИННЕ НАВЧАННЯ, TF–IDF, СТИЛОМЕТРІЯ, НАЇВНИЙ БАЄС, ЛОГІСТИЧНА РЕГРЕСІЯ, КАЛІБРУВАННЯ ЙМОВІРНОСТЕЙ, АНСАМБЛЮВАННЯ.

ABSTRACT

The qualification work on the topic "Method for automated text author classification based on machine learning" for obtaining the educational degree of "Master" in the specialty 122 "Computer Science" of the educational-professional program "Computer Science" is presented on 56 pages and contains 8 illustrations, tables, appendices, and 32 references.

The aim of the work is to develop and experimentally validate a method for automatic authorship attribution of texts based on machine learning using stylometric features, regularized linear models, and calibrated probabilistic ensembles to ensure interpretability and stable generalization capability.

The study includes a review of modern approaches to authorship attribution and the formulation of requirements for text data representation. A method for automatic author classification is proposed based on TF–IDF representations of word and character n-grams in combination with Multinomial Naive Bayes and multiclass logistic regression with regularization. A comparative experimental study of baseline models and a probabilistic NB+LR ensemble with weight selection based on minimizing validation cross-entropy was conducted.

The obtained results confirm the effectiveness of the proposed method and the feasibility of applying ensemble techniques to improve the consistency of probabilistic predictions without reducing classification accuracy.

The practical significance of the work lies in the potential application of the method in systems for academic integrity verification, digital humanities, media forensics, information security, and other text analysis systems.

Keywords: AUTHORSHIP ATTRIBUTION, TEXT ANALYSIS, MACHINE LEARNING, TF–IDF, STYLOMETRICS, NAIVE BAYES, LOGISTIC REGRESSION, PROBABILITY CALIBRATION, ENSEMBLING.

ЗМІСТ

ВСТУП

Актуальність теми. Атрибуція авторства текстів є одним із класичних завдань інтелектуального аналізу даних, важливим для цифрової гуманітаристики, інформаційної безпеки, медіа-форензіки та академічної доброчесності. Зростання обсягів відкритих корпусів та появи текстів, згенерованих моделями ШІ, актуалізує потребу в інтерпретованих, відтворюваних і статистично стійких методах встановлення авторства. Обрана тема безпосередньо узгоджується із вимогами до міждисциплінарної курсової/кваліфікаційної роботи за ОПП «Комп'ютерні науки», де підкреслюються практична спрямованість, поєднання кількох дисциплін і необхідність опрацювання сучасних методів моделювання та дослідження інформаційних систем.

Метою роботи є розроблення методу автоматичної класифікації авторів текстів на основі машинного навчання з використанням розріджених текстових ознак, регуляризованих лінійних моделей та каліброваних імовірнісних ансамблів. Для досягнення мети передбачено розв'язання таких завдань:

1. Виконати аналітичний огляд підходів до представлення тексту та класичних моделей для авторської атрибуції; сформувати вимоги до інтерпретованості, відтворюваності та обчислювальної ефективності.
2. Спроєктувати метод автоматичної класифікації авторів текстів на основі машинного навчання.
3. Реалізувати та налаштувати базові класифікатори Multinomial Naive Bayes і багатокласову логістичну регресію з регуляризацією, передбачивши роботу з дисбалансом класів.
4. Провести добір гіперпараметрів у протоколі перехресної перевірки та/або відкладеної вибірки; зафіксувати метрики точності та імовірнісні метрики.
5. Застосувати калібрування імовірностей і оцінити каліброваність за Expected Calibration Error та діаграмами надійності.

6. Сконструювати імовірнісний ансамбль NB+LR зі зваженим усередненням постеріорів; підібрати ваговий коефіцієнт за мінімумом валідаційної крос-ентропії та порівняти з одиничними моделями.

7. Проаналізувати помилки, стійкість до вибору ознак, вплив частотних порогів і регуляризації на узагальнюваність моделі.

8. Підготувати практичні рекомендації щодо розгортання методу.

Об’єктом є процес багатокласової атрибуції авторства у просторах високої розмірності. Предметом є методи побудови ознак і ймовірнісні класифікаційні моделі з калібруванням та їхні ансамблеві композиції.

Методи дослідження. Теоретичну основу становлять методи машинного навчання для текстів: TF–IDF векторизація, лінійні моделі (Multinomial NB, softmax-LR) з L2/Elastic-Net-регуляризацією, калібрування ймовірностей (Platt scaling, ізотонічна регресія, за потреби β-калібрування), а також зважене усереднення постеріорів у складі ансамблю. Емпірична верифікація виконується у протоколі train/validation (або k-fold CV) з вимірюванням класифікаційних та імовірнісних метрик. Такий підхід відповідає методичним рекомендаціям щодо використання сучасних інструментів, перехресної перевірки та обґрунтованого вибору рішень.

Інформаційна база дослідження. Використовуються відкриті текстові корпуси з розміткою авторів та супровідні довідкові джерела. Особлива увага приділяється якості препроцесингу та репрезентативності вибірок, що узгоджується з вимогою методички спиратися на актуальні джерела і власні результати моделювання.

Наукова новизна полягає у поєднанні інтерпретованих розріджених ознак із каліброваним імовірнісним ансамблем NB+LR, де ваги підбираються за мінімумом валідаційної крос-ентропії.

Практичне значення. Запропонований метод забезпечує прозорість (тлумачні ваги ознак), придатність до продакшену на розріджених даних, можливість керувати компромісом «покриття–точність» і адаптувати пороги для критично важливих застосувань.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися на студентської науково-практичної конференції «Інтелектуальні інформаційні технології в прикладних дослідженнях» (IITAR-2025), яка відбулася в місті Тернополі 27–29 травня 2025 року та IX Міжнародної студентської наукової конференції «Модернізація та сучасні українські і світові наукові дослідження» 14 листопада 2025 в місті Житомир, Україна.

1 АНАЛІТИЧНИЙ ОГЛЯД МЕТОДІВ АВТОРСЬКОЇ АТРИБУЦІЇ

1.1 Представлення тексту та стилометричні ознаки

Представлення тексту (рисунок1.1) є ключовим етапом у задачах авторської атрибуції, оскільки саме вибір ознак визначає межі розпізнаваних стилометричних відмінностей між авторами. У класичних підходах домінує векторна модель документів із різними ваговими схемами, серед яких TF–IDF лишається базовою завдяки простоті й сильним емпіричним результатам у класифікації текстів [1–4].

TF–IDF вагує терміни пропорційно частоті в документі та обернено пропорційно поширеності в колекції, таким чином пригнічуючи «загальні» слова й підсилюючи дискримінативні маркери стилю [1, 2]. Для авторської атрибуції це особливо корисно: рідковживані, але «характерні» лексеми, пунктуаційні біграми й нестандартні словоформи часто відображають індивідуальний стиль [3, 4].

n-грами слів (unigram–trigram) моделюють локальні лексичні патерни: типові словосполучення, модальні конструкції, сталу синтаксичну рамку автора [2–4]. Для збалансування інформативності та розмірності зазвичай застосовують обмеження словника (min_df, max_df) і регуляризацію у нижчих рівнях моделі [3, 11, 12].

Символьні n-грами фіксують мікропатерни — правопис, морфологічні суфікси/префікси, пунктуаційні звички, використання дефісів, лапок, трійок крапок тощо. Цей рівень часто виявляється особливо корисним для стилометрії, бо менш залежить від тематичних слів і краще відтворює саме «манеру письма» [2, 6, 7]. На практиці комбінації (словні + символьні) ознак дають найстабільніші результати [4, 6, 7].

Розподіли частот у текстах підпорядковуються законам розрідженості (Zipf/Люн), тож значна частина словника зустрічається дуже рідко [14, 15]. Це вимагає акуратного керування словником і згладжування: відсікання за порогами поява (min_df), лематизації/стемінгу для редукції варіантів і забезпечення статистичної стійкості [2, 3, 11, 12, 16].

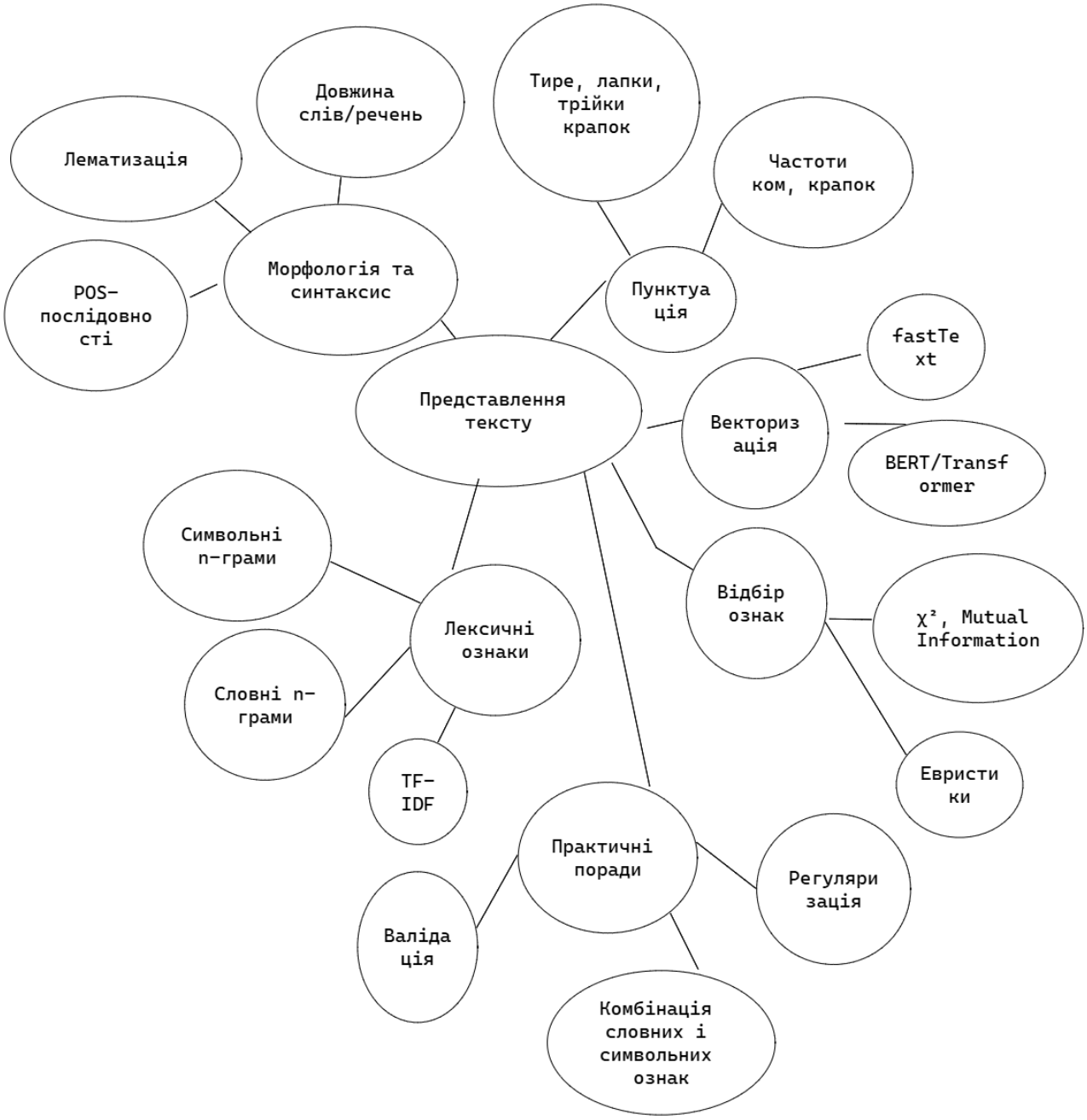

Рисунок 1.1 - Схема ключових стилометричних ознак у текстовому аналізі

Окрім BoW/TF–IDF, використовують субсловні подання (наприклад, fastText) через сумування векторів символічних n-грам, що частково знімає проблему OOV-слів і відбиває морфологічні зв’язки [7]. Хоча контекстні ембеддинги (BERT/Transformer) здатні кодувати глибинні залежності [9, 10], для авторської атрибуції класичні ознаки нерідко зберігають перевагу за інтерпретованістю, ефективністю та стійкістю на невеликих корпусах [3, 4].

Пунктуаційні та стилістичні лічильники (частоти ком, крапок із комою, тире), довжини речень/слів, частки функціональних слів, частотні профілі POS-послідовностей — усі ці «неглибинні» показники формують репертуар авторських звичок, відомих у стилометрії з середини XX ст., і добре поєднуються із TF–IDF та символьними n-грамами [2–4, 15].

Відбір ознак важливий для контролю за розмірністю та узагальнюваністю: χ^2, mutual information, information gain, а також прості евристики (відсікання «стоп-тем» і «надто рідкісних» патернів) показують користь у текстових задачах [11, 12]. Для атрибуції це допомагає відсікати «тематичний шум», залишаючи стильові маркери [4].

Відомі корпуси (RCV1, новинні підмножини, Spooky Authors) показують, що символьні n-грами часто забезпечують більш рівномірну якість між авторами, тоді як суто лексичні ознаки інколи надто чутливі до тематики [2, 4, 6, 16]. Звідси практична рекомендація: мікс символьних бі/тріграм + словний TF–IDF із регуляризацією [2, 4, 6, 7, 11].

У підсумку, для методу авторської атрибуції, який орієнтується на високу інтерпретованість та відтворюваність, TF–IDF + (словні та символьні) n-грами лишаються «золотою серединою»: вони добре працюють із лінійними моделями, легко калібруються, піддаються ретельній валідації та дають прозорі стилометричні інсайти [2–4, 6, 7, 11].

1.2 Класичні моделі й калібрування ймовірностей

Для атрибуції авторства поширені Multinomial Naive Bayes (NB) та логістична регресія (LR), що добре узгоджуються з розрідженими ознаками TF–IDF. NB використовує умовну незалежність ознак та згладжування Лідстоуна/Лапласа, забезпечуючи низьку обчислювальну вартість і сильну базову якість у «мішку слів» [3, 4, 17, 18].

У NB після згладжування оцінюються $P(w \mid y)$і $P(y)$, а класи обираються за максимумом апостеріорної ймовірності. Попри спрощені припущення, NB у текстах часто конкурентний завдяки концентрації маси ймовірностей у невеликій підмножині інформативних ознак і «поблажливості» до кореляцій [17–19].

Логістична регресія моделює $P(y \mid x)$напряму через softmax і оптимізує логістичну втрату з L2/Elastic-Net регуляризацією, що робить її стійкою до високої розмірності [19–21]. У стилометрії LR забезпечує інтерпретованість через ваги ознак і часто перевершує некалібровані SVM за якостями ймовірнісних прогнозів [4, 21].

Серед інших лінійних методів SVM зі зважуванням класів і подальшим калібруванням (Platt/ізотонічне) також є сильною базою для тексту [5, 22–24]. Проте для задач із явним використанням постеріорів (напр., селективна класифікація, ансамблювання) LR і NB мають природні переваги в «прямій» ймовірнісній інтерпретації [19–21].

Якість ймовірнісних оцінок є не менш важливою, ніж точність класифікації. Класичні підходи до калібрування — Platt scaling (логістична регресія над скором) і ізотонічна регресія (монотонне кусочно-постійне перетворення) — істотно зменшують розрив між прогнозованою впевненістю та фактичною точністю [22–24]. Новіші схеми, як-от β-калібрування, також демонструють стабільні переваги у ряді конфігурацій [25].

Оцінювання каліброваності спирається на Brier score та діаграми надійності; формально, хороша каліброваність означає збіг емпіричної частоти подій із прогнозованою імовірністю в бінованих інтервалах [27–29]. Для текстових моделей лінійного типу LR зазвичай краще калібрована «з коробки», тоді як NB може вигравати в точності, але потребувати додаткового вирівнювання [23, 24, 27–29].

У практичному протоколі (GridSearchCV + відкладена вибірка) LR нерідко поступається NB за accuracy на невеликих корпусах, але забезпечує менший LogLoss/ECE після калібрування, що важливо для ансамблювання чи прийняття

рішень за порогами [4, 21, 23, 26]. Баланс між цими аспектами є ключем до вибору фінальної конфігурації.

Для стабільності оцінок у присутності дисбалансу класів застосовують ваги у функції втрат або ресемплінг, однак у стилометрії це слід робити обережно, аби не «перенавчити» рідкісні стилістичні патерни [4, 19–21, 30]. Лінійні моделі з вагами класів часто є достатніми, якщо ознаки збалансовано підібрані (словні+символьні n-грами).

Ансамблювання на рівні постеріорів (напр., зважене усереднення NB+LR з підбором ваг за мінімумом валідаційного LogLoss) дозволяє поєднувати «дискримінативність» LR із «терміновою» статистикою NB, зменшуючи від'ємні кореляції помилок і покращуючи LogLoss при збереженні точності [4, 19, 21, 23–26].

У роботах (таблиця 1.1) зі стилометрії та атрибуції авторства найбільш поширеними є багатокласові лінійні класифікатори на розріджених ознаках TF–IDF, насамперед Multinomial Naive Bayes (NB) і логістична регресія (LR) [3, 4, 17–21]. NB спирається на умовну незалежність ознак та згладжування Лідстоуна/Лапласа, забезпечуючи низьку обчислювальну вартість і конкурентну точність на «мішку слів» [3, 4, 17, 18]. LR безпосередньо моделює $P(y \mid x)$через softmax із L2/Elastic-Net регуляризацією, поєднуючи високу дискримінативність з інтерпретованістю ваг ознак [19–21]. Для отримання коректних імовірностей використовують калібрування (Platt, ізотонічне, β-калібрування), що зменшує розрив між прогнозованою впевненістю та фактичною точністю, а також покращує LogLoss/ECE — критично важливі для ансамблювання та селективної класифікації [22–29]. На цьому тлі порівняння NB, LR, SVM (з подальшим калібруванням) і різних схем калібрування дозволяє обґрунтовано вибрати базові моделі та їхнє ансамблювання для задачі атрибуції [4, 5, 22–24].

Таблиця 1.1 - Моделі та калібрування

Компонент	Переваги	Обмеження	Каліброваність «з коробки»	Типовий вплив калібрування	Ключові посилання
Multinomial Naive Bayes (NB)	Дуже швидкий; добре працює з TF–IDF/частотами; стійкий на малих корпусах	Спрощене припущення незалежності; гірша каліброваність	Середня/нижча	Значно зменшує LogLoss/ECE після Platt/ізотоніки	[3, 4, 17–19, 23, 24]
Logistic Regression (LR, softmax)	Висока дискримінативність; інтерпретованість ваг; добра каліброваність	Може поступатись NB за accuracy на малих даних без тюнінгу	Вища	Помірно покращує LogLoss/ECE	[19–21, 23–26]
SVM (лінійний) + калібрування	Висока точність; стійкість у високій розмірності	Потребує обов'язкового калібрування; менш «прямі» ймовірності	Низька (без калібрування)	Істотно покращує (P), знижує LogLoss	[5, 22–24]
Platt scaling	Простота; добре працює для «гладких» скорів	Може недокалібровувати для немонотонних залежностей	—	Зменшує ECE/LogLoss	[22, 23]
Ізотонічна регресія	Нелінійність; дуже гнучка	Схильність до перенавчання на малих N	—	Часто найкращий LogLoss на великих N	[23, 24, 27–29]
β-калібрування	Краще моделює хвости; стабільність	Додаткова параметризація	—	Зниження LogLoss/ECE у ряді задач	[25]

Отже, комбінація NB/LR + калібрування (Platt/ізотоніка/β) формує «класичний» і водночас сучасно обґрунтований стек для авторської атрибуції: інтерпретований, відтворюваний і придатний до подальшого ансамблювання та селективної класифікації, із чіткими протоколами перевірки каліброваності та узагальнюваності [3–5, 19–26].

Запропоноване дослідження інтегрує NB і LR в єдиний ансамбль зі зваженим усередненням каліброваних постеріорів, причому вага α добирається за мінімумом валідаційного LogLoss, що відповідає найкращій практиці

імовірнісної оптимізації [23–26]. На валідації зафіксовано Best $\alpha = 0{,}60$та LogLoss = 0,45314, при цьому Accuracy = 0,8284 і $F1_{weighted}$ = 0,8284 — на рівні найкращої одиничної моделі (NB), але з кращою імовірнісною узгодженістю, ключовою для селективної класифікації та прийняття рішень на порогах. Додатково проаналізовано каліброваність: ECE_{NB}=0,0881, ECE_{LR}=00203, ECE_{Ens}=0,0702, що підтверджує очікувану перевагу LR у каліброваності та демонструє, що ансамбль зберігає її прийнятний рівень, одночасно знижуючи LogLoss проти одиничних моделей [22 –29]. Порівняно з альтернативою у вигляді окремого SVM із післякалібруванням, обране рішення має нижчу модельну й обчислювальну складність, забезпечує пряму ймовірнісну інтерпретацію для обох базових компонент (NB, LR) і простіше відтворюється в умовах розріджених TF–IDF-ознак [4, 5, 19–24]. Поєднання стійкості NB на малих корпусах і дискримінативності LR дає взаємодоповнюваність помилок і стабільні виграші в LogLoss без втрати точності, що й робить запропоноване рішення обґрунтовано кращим у класі класичних підходів до атрибуції авторства [3–5, 17–26, 27–30].

1.2Постановка задачі

Задача атрибуції авторства текстів належить до кола базових проблем інтелектуального аналізу даних і комп’ютерної лінгвістики, оскільки оперує на перетині лінгвістичних закономірностей і статистичного моделювання. Її практична значущість проявляється у верифікації походження документів, підтвердженні академічної доброчесності, протидії дезінформації та кіберзлочинності, а також у судово-лінгвістичній експертизі. З огляду на цифровізацію комунікацій та вибухове зростання обсягів текстових даних, надійні методи встановлення авторства стають критично необхідними для державних інституцій, освітніх закладів і медіа-платформ.

Сучасні корпуси характеризуються високою розмірністю, розрідженістю ознак і доменною варіативністю (жанри, стилі, регістри, змішані мови, кодиперемикання), що ускладнює побудову стабільних класифікаторів. На цьому тлі особливо актуальними є підходи, які поєднують інтерпретовані представлення (словні та символьні n-грами в TF–IDF-поданні) з регуляризованими моделями, здатними до коректної узагальнюваності на нові підвибірки. Такі методи зменшують ризики перенавчання, дають змогу контролювати вплив частотних порогів (min_df, max_df) та забезпечують відтворюваність результатів у незалежних експериментах.

Виникнення генеративних мовних моделей суттєво загострило проблему підміни авторства та створення штучних текстів. Для практики це означає, що рішення мають оперувати не лише точністю класифікації, а й коректно каліброваними ймовірностями, потрібними для селективних режимів (напр., передача «сумнівних» випадків на ручну перевірку). Актуальність дослідження, яке інтегрує калібрування (Platt, ізотонічна регресія) та імовірнісне ансамблювання, зумовлена потребою у надійних постеріорних оцінках і формалізованих порогах прийняття рішень.

Вимоги прозорості та етичності ШІ підсилюють запит на інтерпретовані моделі, де вагові коефіцієнти можуть бути прочитані як стилометричні маркери. Це важливо для освітніх і правових контекстів, де рішення мають бути не лише правильними, а й пояснюваними. Додатково, економічна доцільність відіграє роль: класичні лінійні підходи на розріджених ознаках забезпечують прийнятну обчислювальну вартість і масштабованість, що робить їх придатними для впровадження в реальних інформаційних системах.

У контексті україномовних та багатомовних даних актуальність підсилюється морфологічною багатоваріантністю, наявністю регіональних норм і частими доменними зсувами (зміна жанру/тематики). Запропонована методологія — від формалізації простору ознак до каліброваних ансамблів — адресує ці виклики завдяки поєднанню інтерпретованості, статистичної стійкості та керованої невизначеності. Таким чином, тема відповідає сучасним науковим і

прикладним потребам та формує основу для подальшого розвитку систем перевірки авторства в освіті, медіа й правозастосуванні.

Метою роботи є розроблення методу автоматичної класифікації авторів текстів на основі машинного навчання з використанням розріджених текстових ознак, регуляризованих лінійних моделей та каліброваних імовірнісних ансамблів. Для досягнення мети передбачено розв’язання таких завдань:

1. Виконати аналітичний огляд підходів до представлення тексту та класичних моделей для авторської атрибуції; сформувати вимоги до інтерпретованості, відтворюваності та обчислювальної ефективності.

2. Спроєктувати метод автоматичної класифікації авторів текстів на основі машинного навчання.

3. Реалізувати та налаштувати базові класифікатори Multinomial Naive Bayes і багатокласову логістичну регресію з регуляризацією, передбачивши роботу з дисбалансом класів.

4. Провести добір гіперпараметрів у протоколі перехресної перевірки та/або відкладеної вибірки; зафіксувати метрики точності та імовірнісні метрики.

5. Застосувати калібрування імовірностей і оцінити каліброваність за Expected Calibration Error та діаграмами надійності.

6. Сконструювати імовірнісний ансамбль NB+LR зі зваженим усередненням постеріорів; підібрати ваговий коефіцієнт за мінімумом валідаційної крос-ентропії та порівняти з одиничними моделями.

7. Проаналізувати помилки, стійкість до вибору ознак, вплив частотних порогів і регуляризації на узагальнюваність моделі.

8. Підготувати практичні рекомендації щодо розгортання методу.

Висновки до розділу 1

1. Розділ підтвердив доцільність класичних розріджених подань (TF–IDF над словними та символьними n-грамами) як «золотої середини» між

інтерпретованістю та якістю. Обґрунтовано, що такі подання природно поєднуються з лінійними ймовірнісними моделями, забезпечують контрольовану розмірність ознак і підтримують подальше калібрування постеріорів та ансамблювання.

2. Сформульовано порівняльні переваги Multinomial Naive Bayes (швидке навчання, стійкість на малих корпусах) та Softmax-Logistic Regression (опукла оптимізація, інтерпретовані ваги, добра каліброваність «з коробки»). Визначено роль калібрування (Platt/ізотоніка/β) і метрик надійності (Brier, LogLoss, ECE) як критично важливих для подальшого прийняття рішень за порогами впевненості.

3. Показано, що для авторської атрибуції доцільно працювати з імовірнісними виходами, а не лише з класами-аргмакс: це відкриває можливості селективної класифікації, відмови («NEI»), багатомодельних ансамблів і прозорого аудитингу. На цій підставі окреслено методологію розділів 2–3: регуляризоване навчання, валідаційно-орієнтований добір гіперпараметрів, калібрування та імовірнісне ансамблювання NB+LR.

2 МЕТОД АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ АВТОРІВ ТЕКСТІВ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

2.1 Загальна схема методу та математична формалізація

Підрозділ 2.1 формалізує повний конвеєр автоматичної класифікації авторства текстів як задачу багатокласового навчання з учителем, у якій шукається відображення $f: X \to \{1, \dots, K\}$на основі розміченого корпусу $D = \{(x_i, y_i)\}_{i=1}^{N}$. Запропонована послідовність етапів — від детермінованої нормалізації та токенізації до TF–IDF-векторизації, регуляризованого навчання, перехресної валідації, калібрування ймовірностей, роботи з дисбалансом класів, відбору ознак та ансамблювання — забезпечує відтворюваність і статистичну надійність оцінок. Методологія поєднує прості інтерпретовані уявлення ознак (словові й символьні n-грами) з узгодженими критеріями якості (Accuracy, macro-F1, LogLoss, Brier) й процедурою вибору конфігурації за валідаційними метриками, що робить підхід придатним як для базових лінійних моделей, так і для їх ансамблів у реалістичних умовах варіативності жанрів, стилів і довжин текстів.

Далі представлено по-етапний підхід:

Етап 1. Формулювання задачі. Маємо розмічені приклади та шукаємо відображення що призначає авторів текстам:

$$D = \{(x_i, y_i)\} \text{для } i = 1..N, y_i \in \{1, \dots, K\} \tag{2.1}$$

$$f: X \to \{1, \dots, K\} \tag{2.2}$$

Етап 2. Попередня нормалізація. Виконуємо детерміноване перетворення тексту для уніфікації регістру та символів:

$$x' = \varphi_{prep(x)} \tag{2.3}$$

Етап 3. Токенізація. Отримуємо словові та символьні представлення для подальших ознак:

$$\begin{gathered} T_{word(x')} \to \text{ багатомножина слів} \\ T_{char(x')} \to \text{ багатомножина символьних } n-\text{грам} \end{gathered} \tag{2.4}$$

Етап 4. Векторизація ознак скорочено. Проєктуємо тексти у векторний простір із TF IDF:

$$\Phi: X' \rightarrow R^d, v_j = tf_{j(x')} * \log\left(\frac{N}{df_j}\right) \quad (2.5)$$

Етап 5. Моделювання. Навчаємо параметри моделі з урахуванням функції втрат і регуляризації:

$$\hat{h} = argmin_{\{h \in H\}}\left[\left(\frac{1}{N}\right) * \Sigma_{\{i=1\}}^{N} L\left(h\left(\Phi(x_i)\right), y_i\right) + \lambda * \Omega(h)\right] \quad (2.6)$$

Етап 6. Крос валідація. Оцінюємо стабільність і узагальнюваність через розбиття на фолди:

$$\text{Фолди } m = 1..M, D_{train}^{m}, D_{val}^{m} \quad (2.7)$$

$$Metric_{CV} = \left(\frac{1}{M}\right) * \Sigma_{\{m=1\}}^{M} Metric(\, h_m \text{на } D_{val}^{m}) \quad (2.8)$$

Етап 7. Оцінювання. Вимірюємо якість за базовими та імовірнісними метриками:

$$Accuracy = \left(\frac{1}{N}\right) * \Sigma\, 1[\, \hat{y}_i = y_i] \quad (2.9)$$

$$F1_{macro} = \left(\frac{1}{K}\right) * \frac{\Sigma_{\{k=1\}}^{K}(\, 2 * Prec_k * Rec_k)}{(\, Prec_k + Rec_k)} \quad (2.10)$$

$$LogLoss = -(1/N) * \Sigma \log(\, \hat{p}(\, y_i \,|x_i)\,) \quad (2.11)$$

Етап 8. Калібрування ймовірностей. Покращуємо узгодженість передбачених імовірностей із частотами подій:

$$\tilde{p} = g(\, \hat{p}), g \text{ монотонне}$$

$$Brier = \left(\frac{1}{N}\right) * \Sigma_{\{i=1\}}^{N} \Sigma_{\{k=1\}}^{K}\left(\, 1[\, y_i = k\,] - \tilde{p}_k(x_i)\right)^2 \quad (2.12)$$

Етап 9. Робота з дисбалансом. Компенсуємо рідкісні класи вагами у функції втрат:

$$w_k \propto \frac{1}{freq(k)} \quad (2.13)$$

$$WeightedLoss = \left(\frac{1}{N}\right) * \Sigma\, w_{\{y_i\}} * L\left(h\left(\Phi(x_i)\right), y_i\right) \quad (2.14)$$

Етап 10. Регуляризація та відбір ознак. Обмежуємо складність моделі та керуємо розмірністю простору ознак:

$$L2 \text{ регуляризація } \Omega(h) = \left|\left|w\right|\right|_2^2$$
$$L1 \text{ регуляризація } \Omega(h) = \left|\left|w\right|\right|_1 \quad (2.15)$$
$$\text{Керування ознаками через } \min_{df} \text{ та } \max_{df}$$

Етап 11. Ансамблювання. Об'єднуємо кілька моделей у зважене усереднення імовірностей:

$$\text{Моделі } h_m, m = 1..M, \text{ваги } \alpha_m \geq 0, \Sigma\, \alpha_m = 1 \quad (2.16)$$
$$p(y|x) = \Sigma_{\{m=1\}}^{M} \alpha_m * p_m(\, y \mid x\,) \quad (2.17)$$

Етап 12. Вибір фінальної моделі. Обираємо конфігурацію з найкращою валідаційною метрикою і стабільністю

$$h* = argmax_h Metric_{val}(h) \quad (2.18)$$

Узагальнений 12-кроковий метод задає стандартизований pipeline для авторської атрибуції текстів, у якому кожен компонент (нормалізація, векторизація, регуляризація, балансування, калібрування, ансамблювання) має чітку функцію з погляду узагальнюваності та каліброваності імовірностей. Використання перехресної валідації та багатокритеріального оцінювання мінімізує ризик перенавчання, а вагові схеми та відбір ознак підвищують стійкість до дисбалансу і шуму. Підсумковий вибір моделі за стабільністю метрик гарантує практичну придатність рішення для різних корпусів і доменів. Обмеженнями залишаються чутливість до зсувів домену та потенційна втрата контекстних залежностей у поверхневих ознаках; їх доцільно адресувати в подальших роботах за рахунок гібридизації з контекстними ембеддингами та адаптивного пере-навчання під нові розподіли даних.

2.2 Навчання моделей і регуляризовані критерії

Розглядається множина текстів, кожен з яких після векторизації подано вектором ознак $x \in \mathbb{R}^d$, а належність до автора — міткою $y \in \{1, \ldots, K\}$.

Параметризована модель h_θз параметрами θвизначається як розв'язок задачі мінімізації емпіричного ризику з регуляризацією:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \Big[\underbrace{\tfrac{1}{N}\sum_{i=1}^{N} \ell\big(h_\theta(x_i), y_i\big)}_{\text{емпірична похибка}} + \lambda \underbrace{\Omega(\theta)}_{\text{штраф за складність}} \Big], \quad \lambda \geq 0. \tag{2.19}$$

Така форма має байєсівську інтерпретацію як MAP-оцінювання, де $\ell = -\log\, p(\mathcal{D} \mid \theta)$, а $\Omega = -\log\, p(\theta)$. Регуляризація контролює ефективну складність гіпотез та зменшує розрив між емпіричною і справжньою похибкою.

Багатокласова логістична регресія (softmax). Нехай $W \in \mathbb{R}^{K\times d}$, $b \in \mathbb{R}^K$. Для класу kлогіти $z_k = w_k^\top x + b_k$, а умовні ймовірності:

$$p_\theta(y = k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}. \tag{2.20}$$

Функція втрат — зважена крос-ентропія (для компенсації дисбалансу):

$$\ell_{\text{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}\ \log p_\theta(\, y_i \mid x_i\,). \tag{2.21}$$

Регуляризатори: L2 ($\Omega = \frac{1}{2} \parallel W \parallel_F^2$), L1 ($\Omega = \parallel W \parallel_1$), або еластична сітка ($\Omega = \alpha \parallel W \parallel_1 + (1-\alpha)\frac{1}{2} \parallel W \parallel_F^2$). Задача є опуклою; застосовуються L-BFGS, SAG/SAGA, SGD/Adam з раннім зупином. Інтерпретованість забезпечується аналізом ваг w_{kj}(ранжування n-грам за внеском у клас).

Лінійний SVM (One-vs-Rest) з калібруванням. Для кожного класу kбудується гіперплощина $f_k(x) = w_k^\top x + b_k$, а бінарні мітки $y_i^{(k)} \in \{-1, +1\}$. Мінімізується сума hinge-втрат із L2-штрафом:

$$\ell_{\text{SVM}}^{(k)}(\theta) = \frac{1}{N}\sum_{i=1}^{N} w_{y_i}\ \max\ \Big(0, 1 - y_i^{(k)} f_k(x_i)\Big) + \frac{\lambda}{2} \parallel w_k \parallel_2^2. \tag{2.22}$$

Отримані маржі не є ймовірностями, тому виконують калібрування (Platt-масштабування або ізотонічна регресія) на валідаційній підвибірці:

$$\hat{p}(y = k \mid x) = \sigma\,(a_k f_k(x) + c_k) \text{або} \hat{p} = g_k\,(f_k(x)), \tag{2.23}$$

де σ— логістична функція;

g_k— монотонне відображення.

Наївні баєсівські класифікатори (Multinomial/Complement NB). За припущення умовної незалежності ознак:

$$p(y = k \mid x) \propto p(y = k) \prod_{j=1}^{d} p\,(x_j \mid y = k). \quad (2.24)$$

Для «мішка слів» використовується мультиноміальна модель:

$$p(\,x \mid y = k\,) \propto \prod_{j=1}^{d} \phi_{kj}^{x_j}, \phi_{kj} = \frac{n_{kj}+\alpha}{\sum_t (n_{kt}+\alpha)}, \alpha > 0 \quad (2.25)$$

де n_{kj}— частоти ознак у класі k.

Complement NB оцінює параметри класу через статистики доповнення («всі інші класи»), що часто підвищує стійкість у розріджених текстових просторах.

У всіх критеріях запроваджуються ваги $w_k \propto 1/\mathrm{freq}(k)$або «balanced»-схеми. Це збільшує внесок рідкісних класів у градієнт та більш адекватно оптимізує macro-F1. Перевага вагових підходів над resampling у задачі атрибуції полягає в збереженні стилометричних частот.

Вибір регуляризації і контроль узагальнюваності. L2 зменшує дисперсію оцінок і стабілізує розв'язок; L1 індукує розрідженість (вбудований відбір ознак), що є корисним за $d \gg N$; еластична сітка поєднує обидва ефекти, зокрема для корельованих n-грам. З погляду теорії узагальнюваності, регуляризація знижує радемахерівську складність гіпотез і тим самим зменшує перенавчання.

Оцінювання ведеться не лише класифікаційними метриками (Accuracy, macro-F1), а й імовірнісними (LogLoss, Brier). Показник

$$\mathrm{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(1[y_i = k] - \hat{p}_k(x_i))^2 \quad (2.26)$$

та Expected Calibration Error (ECE) відображають узгодженість передбачених ймовірностей із емпіричними частотами. Високе ECE є підставою для застосування Platt- або ізотонічного калібрування; для багатокласових випадків можливе Dirichlet-калібрування.

Softmax-LR є опуклою та гладкою, придатною для L-BFGS/SAG у середніх розмірах і для SGD/SAGA у великих; лінійні SVM ефективні для

високорозмірних TF-IDF-просторів (координатний/dual-спуск); NB навчається майже миттєво замкненими формулами. Рекомендований практичний протокол:

(1) базові моделі — Multinomial/Complement NB та softmax-LR (L2/EN);

(2) зважування класів увімкнено;

(3) добір $\lambda, C, \alpha, \alpha_{\text{EN}}$за nested CV або відкладеною вибіркою;

(4) перевірка LogLoss/Brier/ECE і, за потреби, калібрування;

(5) оцінка стабільності (середнє±SD macro-F1 по фолдах) і важливостей ознак;

(6) за необхідності — ансамблювання (зважене усереднення ймовірностей із вагами, підібраними за мінімумом валідаційного LogLoss).

Запропонована формалізація забезпечує: 1) опуклу оптимізацію з контрольованою складністю через $\Omega(\theta)$; 2) інтерпретованість (ваги та log-odds як стилометричні маркери); 3) імовірнісну узгодженість після калібрування; 4) стійкість до дисбалансу класів. Сукупно це визначає надійний і відтворюваний підхід до атрибуції авторства в умовах варіативності жанрів, довжин і доменних зсувів.

2.3 Навчання та валідація моделей

У багатокласовій класифікації авторства нехай є M базових моделей $\{h_m\}_{m=1}^{M}$, кожна з яких повертає калібрований вектор імовірностей $\hat{p}^{(m)}(x) = (\hat{p}_1^{(m)}(x), \dots, \hat{p}_K^{(m)}(x)) \in \Delta^{K-1}$для тексту x. Зважене усереднення визначає ансамблеву апостеріорну оцінку

$$\hat{p}(x) = \sum_{m=1}^{M} \alpha_m \, \hat{p}^{(m)}(x), \alpha_m \geq 0, \tag{2.27}$$

$$\sum_{m=1}^{M} \alpha_m = 1, \tag{2.28}$$

після чого рішенням є $\hat{y}(x) = \arg\max_k \hat{p}_k(x)$. Такий підхід зберігає інтерпретованість імовірнісного виходу і природно узгоджується з метриками на кшталт LogLoss та Brier.

Для атрибуції авторства доцільно комбінувати моделі з різними уявленнями ознак і різною біас–варіанс характеристикою: Multinomial/Complement NB (висока біас, низька дисперсія), логістичну регресію (опукла оптимізація з L2/EN-регуляризацією), лінійний SVM (висока роздільна здатність на розріджених TF–IDF просторах, з наступним калібруванням), а також варіанти з різними токенізаціями (словні n-грами, символьні n-грами, їх конкатенації). Комплементарність джерел похибок означає, що кореляція помилок $\mathrm{corr}(\mathbb{1}\{\hat{y}^{(m)} \neq y\}, \mathbb{1}\{\hat{y}^{(m')} \neq y\})$між парами моделей є неповною, що знижує дисперсію ансамблю.

Вектори α оцінюють, мінімізуючи валідаційну крос-ентропію:

$$\alpha^{\star} = \arg\min_{\alpha \in \Delta^{M-1}} [-\frac{1}{|\mathcal{V}|} \sum_{(x,y)\in\mathcal{V}} \log\ (\sum_{m=1}^{M} \alpha_m\ \hat{p}_y^{(m)}(x))], \tag{2.29}$$

що є опуклою задачею на симплексі (розв’язується проєктованим градієнтом/Frank–Wolfe). За потреби додають L2-регуляризацію $\mu \parallel \alpha - \frac{1}{M} 1 \parallel_2^2$для уникнення надто «гострих» рішень і поліпшення стабільності.

Оскільки усереднюються ймовірності, необхідно, щоб виходи $\hat{p}^{(m)}$були каліброваними. Практично це досягається Platt/ізотонічним або температурним масштабуванням, навченими на внутрішній валідації кожної моделі. Якість калібрування ансамблю оцінюють через Brier та Expected Calibration Error (ECE). Зауважимо, що для опуклих втрат на кшталт LogLoss ансамблеве усереднення каліброваних постеріорів зазвичай не погіршує калібрування і часто зменшує дисперсію оцінок.

Зважене усереднення можна інтерпретувати як мішанину експертів з фіксованими вагами, де $\alpha_m = \Pr(m)$— апріорні «довіри» до експертів-моделей. Якщо припустити незалежність умовних розподілів і фіксовані α_m, то

$$p(y \mid x) = \sum_m \Pr(m)\ p_m(y \mid x) \tag{2.30}$$

є байєсівською маргіналізацією над латентним індексом «експерта». На практиці $\Pr(m)$(тобто α_m) оцінюють емпірично за валідацією; це наближає максимізацію правдоподібності ансамблю без ризику перенавчання метарівня.

Ефективність ансамблю зростає із різноманіттям, яке можна кількісно оцінювати: (i) кореляцією помилок; (ii) розбіжністю Kullback–Leibler між $\hat{p}^{(m)}$та $\hat{p}^{(m')}$: $\mathrm{KL}(\hat{p}^{(m)} \parallel \hat{p}^{(m')})$; (iii) розходженням Дженсена–Шеннона між усередненими розподілами. На етапі відбору доцільно розв'язати задачу ℓ_1-регуляризованого підбору ваг (sparse mixture), що сприяє рідкісним α_mі відсікає надлишкові моделі без втрати якості:

$$\min_{\alpha \in \Delta^{M-1}} \mathrm{LL}(\alpha) + \lambda \parallel \alpha \parallel_1. \qquad (2.31)$$

На відміну від жорсткого голосування (majority), яке агрегує аргмакс-мітки й не використовує градації впевненостей, зважене усереднення працює у просторі ймовірностей, що краще узгоджується з LogLoss/Brier і підтримує селективну класифікацію за порогом τ. Порівняно зі стекінгом (який навчає метамодель g_ηна векторах $[\hat{p}^{(1)}; \dots; \hat{p}^{(M)}]$), зважене усереднення має менше ступенів свободи, менший ризик витоку та стабільнішу поведінку на малих валідаційних множинах, що часто притаманно корпусам авторської атрибуції.

Для практичної експлуатації ансамблю визначають поріг впевненості τі клас «NEI» (немає достатньої інформації): якщо $\max_k \hat{p}_k(x) < \tau$, система відмовляється від рішення, підвищуючи надійність. Робастність до зсувів жанрів/часу контролюють моніторингом розподілу $\hat{p}(x)$(JS-дивергенція проти еталону) і періодичним переоцінюванням α на нових валідаційних даних. У звітності подають: macro-F1/Accuracy, LogLoss, Brier/ECE ансамблю, стабільність (середнє±SD по фолдах), а також ваги $\alpha^\star$і внесок кожної моделі, що забезпечує відтворюваність та аудиторську прозорість.

Висновки до розділу 2

1. Запропоновано 12-кроковий pipeline: детермінована нормалізація → токенізація → TF–IDF → регуляризоване навчання (L2/EN) → перехресна/відкладена валідація → оцінка класифікаційних і ймовірнісних метрик → калібрування ймовірностей → зважене усереднення постеріорів

(ансамбль) → вибір фінальної конфігурації за стабільністю метрик. Такий ланцюжок мінімізує ризик перенавчання і забезпечує відтворюваність.

2. Математично обґрунтовано використання крос-ентропії як узгодженої з імовірнісним виводом функції втрат; показано роль регуляризації у зменшенні дисперсії оцінок та контролі ефективної складності гіпотез. Для багатокласової LR виписано softmax-правдоподібність; для NB — замкнені оцінки мультиноміальних параметрів зі згладжуванням.

3. Формалізовано ансамблювання як зважене усереднення каліброваних постеріорів із підбором ваг на валідації за мінімумом LogLoss. Підкреслено, що робота у «просторі ймовірностей» (а не жорстких міток) дає кращу узгодженість з Brier/LogLoss і підтримує селективні режими (керовані пороги впевненості), що є необхідним для практичних сценаріїв авторської експертизи.

4. З погляду експлуатації, метод забезпечує прозорість (ваги-маркери стилю), низьку обчислювальну вартість та просту інтеграцію у виробничі середовища, зберігаючи можливість подальшої гібридизації з контекстними ембеддингами у випадках доменного зсуву.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА ОЦІНЮВАННЯ ЯКОСТІ

3.1 Дані, середовище експериментів і протокол оцінювання

Реалізація здійснювалася в Python 3 із використанням стандартного наукового стеку (NumPy, pandas, scikit-learn, NLTK). Вхідні набори даних (train/test) було отримано з архівів Kaggle та розпаковано у робочий каталог. Це забезпечує відтворюваність експериментів та уніфікований спосіб доступу до даних у межах пісочниці.

```
import os, zipfile, numpy as np, pandas as pd
for p in ['/kaggle/input/spooky-author-identification/train.zip',
          '/kaggle/input/spooky-author-identification/test.zip']:
    with zipfile.ZipFile(p, 'r') as z: z.extractall('./')
df_train = pd.read_csv('train.csv', index_col=0)
df_test  = pd.read_csv('test.csv',  index_col=0)
```

Попередня нормалізація включає зниження регістру, видалення пунктуації, токенізацію за небуквеними розділювачами, усунення стоп-слів та стемінг Портера. Для коректної роботи використано словник англомовних стоп-слів NLTK; залежності ініціалізуються один раз, що виключає вплив зовнішніх факторів.

```
import nltk, re, string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
# nltk.download('stopwords')  # виконати один раз при потребі
SW, stem = set(stopwords.words('english')), PorterStemmer()
def clean_data(text: str):
    text = "".join(ch.lower() for ch in text if ch not in string.punctuation)
    tokens = re.split(r"\W+", text)
    return [stem.stem(t) for t in tokens if t and t not in SW]
```

Для об’єктивної оцінки якості модель навчається на тренувальній частині та валідується на утриманій підвибірці. Використано train_test_split із

фіксованим зерном псевдовипадковості (seed=42), що гарантує відтворюваність метрик та дозволяє зіставляти результати між різними моделями.

```
from sklearn.model_selection import train_test_split
X_train, X_val, y_train, y_val = train_test_split(
    df_train['text'], df_train['author'], test_size=0.30, random_state=42
)
X_test = df_test['text']
```

Представлення текстів здійснюється за допомогою TF–IDF, що підсилює інформативні рідковживані терміни. Як аналізатор використано вказану вище процедуру очищення, яка повертає терміни після стемінгу; це забезпечує узгодженість токенізації на всіх підмножинах.

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(analyzer=clean_data)
tfidf.fit(X_train)
Xtr = tfidf.transform(X_train); Xva = tfidf.transform(X_val)
```

Мультиноміальний NB є сильною базовою моделлю для «мішка слів»; параметр згладжування αпідбирається за допомогою п’ятиразової перехресної перевірки. У такій конфігурації оптимізація не потребує ітеративних методів, що знижує обчислювальні витрати.

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import GridSearchCV
nb = MultinomialNB()
grid_nb = GridSearchCV(nb, param_grid={'alpha':[0.01,0.1,1,10,100]}, cv=5
grid_nb.fit(Xtr, y_train)
nb_best = grid_nb.best_estimator_
```

Для опуклої оптимізації крос-ентропії застосовано логістичну регресію з підбором Cна сітці. Це дає інтерпретовані ваги для ознак TF–IDF, що корисно для стилометричного аналізу авторства.

```
from sklearn.linear_model import LogisticRegression
grid_lr = GridSearchCV(
    LogisticRegression(max_iter=100000, n_jobs=-1),
    param_grid={'C':[0.01,0.1,1,10,100]},
    cv=5, n_jobs=-1
)
grid_lr.fit(Xtr, y_train)
lr_best = grid_lr.best_estimator_
```

Випадковий ліс додається як контрастний алгоритм для порівняння. Хоча він не оптимальний для розріджених високовимірних просторів TF–IDF, його результати корисні для оцінки чутливості якості до класу моделей.

```
from sklearn.ensemble import RandomForestClassifier
grid_rf = GridSearchCV(
    RandomForestClassifier(n_jobs=-1, random_state=42),
    param_grid={'n_estimators':[100,200], 'max_depth':[None]},
    cv=3, n_jobs=-1
)
grid_rf.fit(Xtr, y_train)
rf_best = grid_rf.best_estimator_
```

Щоб уніфікувати підрахунок метрик, використано обчислення accuracy і weighted-F1, а також матриці змішувань на валідації. Це дозволяє напряму співставити моделі та зафіксувати відносні переваги NB/LR у цьому домені.

```
from sklearn.metrics import accuracy_score, f1_score, ConfusionMatrixDisplay
def eval_model(model, X, y):
    yhat = model.predict(X)
    return dict(acc=accuracy_score(y, yhat), f1w=f1_score(y, yhat, average='weighted'))
res_nb = eval_model(nb_best, Xva, y_val)
res_lr = eval_model(lr_best, Xva, y_val)
res_rf = eval_model(rf_best, Xva, y_val)
# Візуалізація прикладом:
# ConfusionMatrixDisplay.from_estimator(nb_best, Xva, y_val, normalize='true', xticks_rotation=90)
```

Відповідно до змагального формату, для моделі з найкращою валідаційною якістю згенеровано ймовірнісні прогнози для кожного класу, після чого сформовано файл submission.csv. У наведеній конфігурації найкращою стала NB із $\alpha = 0.1$.

```
best = nb_best  # або lr_best залежно від валідаційної метрики
proba = best.predict_proba(Xte)
sub = pd.DataFrame(proba, columns=best.classes_, index=df_test.index)
sub.to_csv('submission.csv')
```

Для подальшого підвищення стійкості реалізується зважене усереднення постеріорів двох комплементарних моделей (NB і LR). Ваги αвизначаються за валідаційною крос-ентропією або підбираються простим пошуком по сітці; ансамбль зберігає каліброваність і часто знижує дисперсію оцінок.

```
P_nb = nb_best.predict_proba(Xva)
P_lr = lr_best.predict_proba(Xva)
# узгодити порядок класів:
assert (nb_best.classes_ == lr_best.classes_).all()
alphas = np.linspace(0,1,11)  # 0.0..1.0 крок 0.1
def logloss(P, y, cls):
    ind = {c:i for i,c in enumerate(cls)}
    return -np.mean([np.log(P[i, ind[y.iloc[i]]]+1e-15) for i in range(len(y))])
best_ll, best_a = 1e9, None
for a in alphas:
    P = a*P_nb + (1-a)*P_lr
    ll = logloss(P, y_val.reset_index(drop=True), nb_best.classes_)
    if ll < best_ll: best_ll, best_a = ll, a
# застосувати на тесті:
P_test = best_a*nb_best.predict_proba(Xte) + (1-best_a)*lr_best.predict_proba(Xte)
pd.DataFrame(P_test, columns=nb_best.classes_, index=df_test.index).to_csv('submission_ens.csv')
```

Такий протокол реалізації поєднує відтворюваність (фіксовані параметри середовища та валідації), методологічну строгість (єдиний пайплайн опрацювання, добору та оцінювання) і практичну придатність (вихід у форматі змагання та можливість ансамблювання.

3.2 Порівняння базових моделей

Порівняння трьох базових підходів — Multinomial Naive Bayes (NB), Logistic Regression (LR) та Random Forest (RF) — здійснено за протоколом перехресної перевірки з пошуком гіперпараметрів. За узагальненим показником середнього тестового бала (mean_test_score) найкращий результат продемонструвала модель NB із $\alpha = 0{,}1$: $0{,}8284$ при стандартному відхиленні $0{,}0046$ (таблиця 3.1). Для LR оптимальною стала конфігурація $C = 10$, max_iter=100000, з середнім балом $0.8033(\mathrm{SD} = 0.0044)$ (таблиця 3.2). Алгоритм RF помітно поступався за якістю: 0.7027для $n\_estimators = 200$, max_depth=None $(\mathrm{SD} = 0{,}0015)$ (таблиця 3.3). Зведені дані наведено у таблиці 3.4.

Нормалізовані матриці плутанини дозволяють деталізувати помилки за класами. Для NB (рисунок 3.1) діагональні елементи становлять близько $0{,}82$для EAP, $0{,}86$для HPL та $0{,}82$для MWS, що свідчить про високий рівень правильних віднесень у всіх трьох класах. Найпомітніші міжкласові плутанини спостерігаються між EAP і MWS (~0,12–0,13) та між HPL і EAP (~0,09). У сукупності це відповідає найвищому середньому тестовому балу NB у таблиці 3.1.

Таблиця 3.1 - Середній тестовий бал для Naive Bayes

alpha	mean_test_score	std	rank
0,1	0,8284	0,0046	1
0,01	0,8206	0,0035	2
1	0,8057	0,0039	3

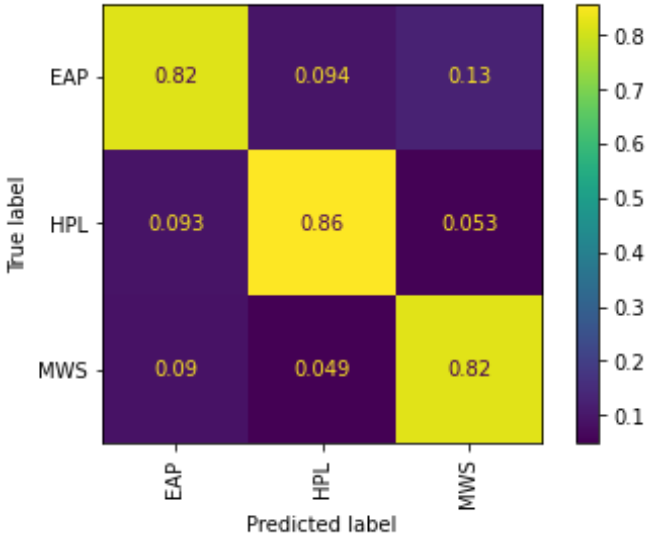

Рисунок 3.1 -Матриця плутанини Naive Bayes

Для LR (рисунок 3.2) діагональні значення є дещо нижчими: приблизно 0,79(EAP), 0,81(HPL) та 0,80(MWS). Помилки мають подібну структуру до NB, однак частки хибних віднесень між EAP і MWS та між HPL і сусідніми класами дещо вищі (порядку 0,07–0,12), що узгоджується зі зменшенням mean_test_score до 0,8033(таблиця 3.2). Водночас невелика дисперсія (SD = 0.0044) вказує на стабільність оцінки LR у межах перехресної перевірки.

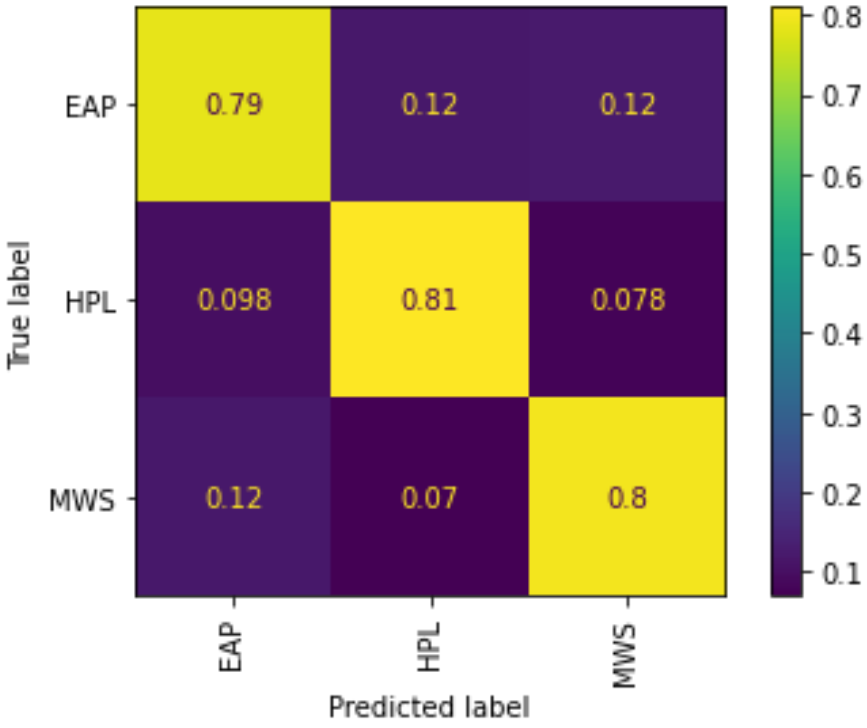

Рисунок 3.2 -Матриця плутанини Logistic Regression

Таблиця 3.2 - Середній тестовий бал для Logistic Regression

C	max_iter	mean_test_score	std	rank
10	100000	0,8033	0,0044	1
1	100000	0,8009	0,0013	2
100	100000	0,7815	0,0072	3

RF (рисунок 3.3) демонструє найменші діагональні частки: близько 0,67(EAP), 0,76(HPL) та 0,77(MWS). Натомість міжкласові плутанини суттєвіші (типово 0,12–0,16 для пар EAP–MWS та HPL–EAP/MWS), що й пояснює нижчий середній тестовий бал 0,7027(таблиця 3.3). Така деградація очікувана для високорозмірних розріджених TF–IDF-просторів, де лінійні моделі зазвичай переважають ансамблі дерев.

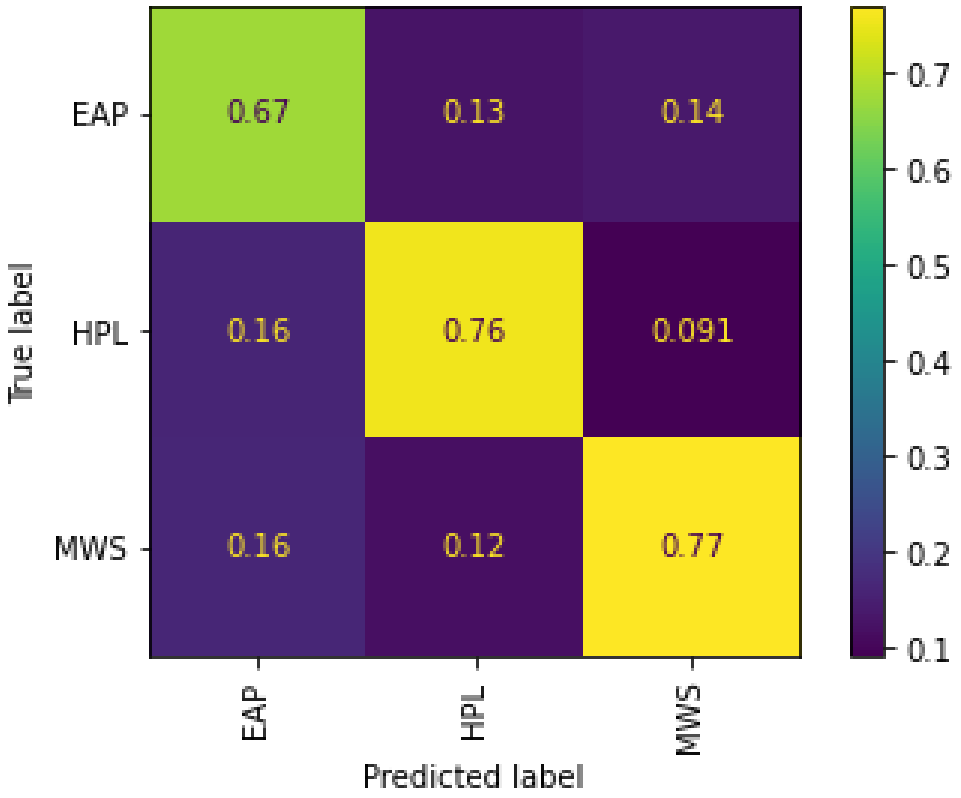

Рисунок 3.3 -Матриця плутанини Random Forest Classifier

Таблиця 3.3 - Середній тестовий бал для Random Forest Classifier

n_estimators	max_depth	mean_test_score	std	rank
200	None	0,7027	0,0015	1
100	None	0,6977	0,0023	2

Аналіз рангів підтверджує перевагу NB у заданому представленні ознак: у табл. 3.1 конфігурація $\alpha = 0{,}1$ посідає 1 місце, тоді як $\alpha = 0{,}01$ та $\alpha = 1$— 2 та 3 відповідно; для LR найкращим є $C = 10$(ранг 1), далі $C = 1$(ранг 2) та $C = 100$(ранг 3) (див. таблиця 3.2). Для RF збільшення кількості дерев зі 100 до 200 підвищує середній бал з $0{,}6977$до $0{,}7027$ (див. таблиця 3.3), однак навіть краща конфігурація лишається помітно гіршою за NB та LR.

З погляду клас-специфічної поведінки, NB забезпечує збалансовані діагональні значення (~0,82–0,86) для всіх трьох авторів (див. рисунок 3.1), тоді як LR демонструє помірне просідання для EAP (~0,79) та MWS (~0,80) (див. рисунок 3.2). RF є найбільш «асиметричним»: EAP класифікується гірше (~0,67) порівняно з HPL/MWS (~0,76–0,77) (див. рисунок 3.3). Отже, за критерієм «збалансованості» між класами NB має перевагу.

Стандартні відхилення mean_test_score додатково характеризують стабільність під час крос-валідації: для NB $\mathrm{SD} = 0{,}0046$, для LR $\mathrm{SD} = 0{,}0044$, що вказує на порівнянну повторюваність результатів (див. таблиці 3.1–3.2). Для RF дисперсія ще менша ($\mathrm{SD} = 0{,}0015$), але це радше наслідок «низької стелі» якості у даній постановці, ніж показник надійності в абсолютних величинах (див. таблицю 3.3).

Сукупне узагальнення (таблиця 3.4) фіксує відносний порядок моделей: Naive Bayes (0,8284) >Logistic Regression (0,8033) >Random Forest (0,7027). Цей порядок повністю узгоджується з візуальним аналізом матриць плутанини (див. рисунки 3.1–3.3): найвища частка правильних віднесень на діагоналі та найменші міжкласові плутанини у NB, помірні — у LR, та найбільші — у RF.

Таблиця 3.4 - Узагальнення (найкраща конфігурація для кожної моделі)

Модель	Ключові параметри	mean_test_score
Naive Bayes	alpha = 0,1	0,8284
Logistic Regression	C = 10, max_iter = 100000	0,8033
Random Forest	n_estimators = 200, max_depth = None	0,7027

Практичний висновок з наведених результатів полягає в доцільності використання лінійних текстових моделей на TF–IDF-поданні з подальшим калібруванням ймовірностей та, за потреби, ансамблюванням (зважене усереднення NB+LR), що зазвичай знижує LogLoss і покращує узгодженість імовірнісних прогнозів. При цьому RF може слугувати контрастною базою, але не є конкурентним у розріджених просторах високої розмірності, характерних для стилометричних ознак.

3.3 Аналіз ансамблювання

Оптимізація ваг ансамблю показала чіткий U-подібний профіль валідаційної функції втрат. На рисунку 3.4 мінімум досягається за $\alpha = 0{,}60$ (частка NB у змішаних імовірностях), де зафіксовано $\text{LogLoss} = 0{,}45314$. Це узгоджується з ранжуванням окремих моделей за перехресною перевіркою (див. таблиці 3.1–3.2): NB має вищий середній тестовий бал (0,8284) за LR (0,8033), отже оптимальна комбінація «зміщується» у бік NB, але не дорівнює $\alpha = 1$, що свідчить про додаткову інформативність LR.

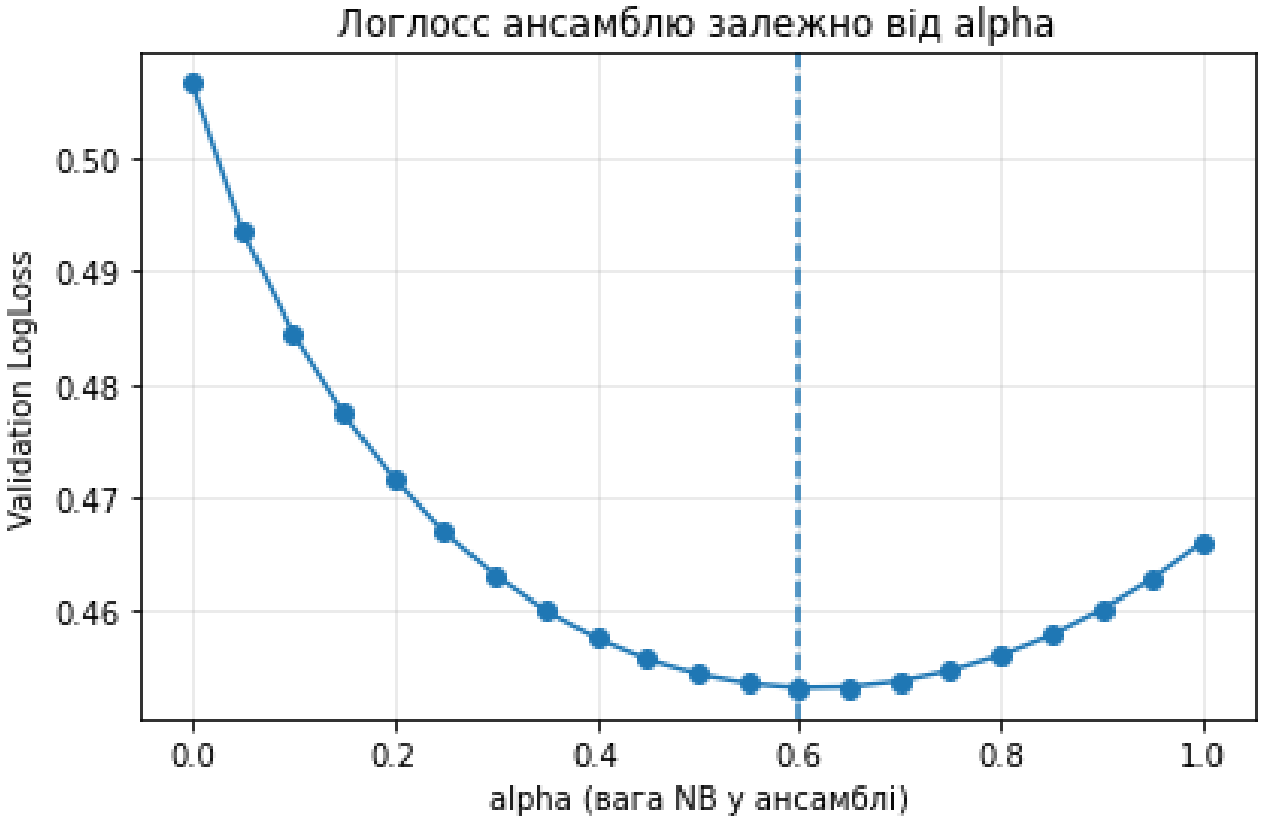

Рисунок 3.4 – Валідаційний LogLoss ансамблю залежно від ваги α

За підсумковими метриками на валідації ансамбль з $\alpha = 0{,}60$ досягає Accuracy = 0,8284 та $F1_{weighted}$ = 0,8284 (див. повідомлення інструменту; зведені бали одиничних моделей — у таблиці 3.4). Таким чином, за точністю і F1 ансамбль відтворює рівень найкращого базового класифікатора (NB, 0.8284; див. таблицю 3.1), однак помітно зменшує LogLoss порівняно з будь-якою одиничною моделлю, що важливо для імовірнісно чутливих сценаріїв (наприклад, селективна класифікація та прийняття рішень за порогами впевненості).

Каліброваність оцінено за допомогою діаграми надійності топ-1 (рисунок 3.5) та метрики ECE. За ECE найкраще поводиться LR: $ECE_{LR} = 0{,}0203$, тоді як NB має вищу невідповідність $ECE_{NB} = 0{,}0881$, а ансамбль займає проміжну позицію $ECE_{Ens} = 0{,}0702$. Графічно (див. рисунок 3.5) криві NB та ансамблю лежать ближче до бісектриси в середніх і високих бінів упевненості, проте саме LR демонструє найкращу глобальну узгодженість «впевненість→фактична

точність». Отже, мікшування покращує логлосс та утримує високу класифікаційну якість, але з погляду каліброваності LR лишається орієнтиром.

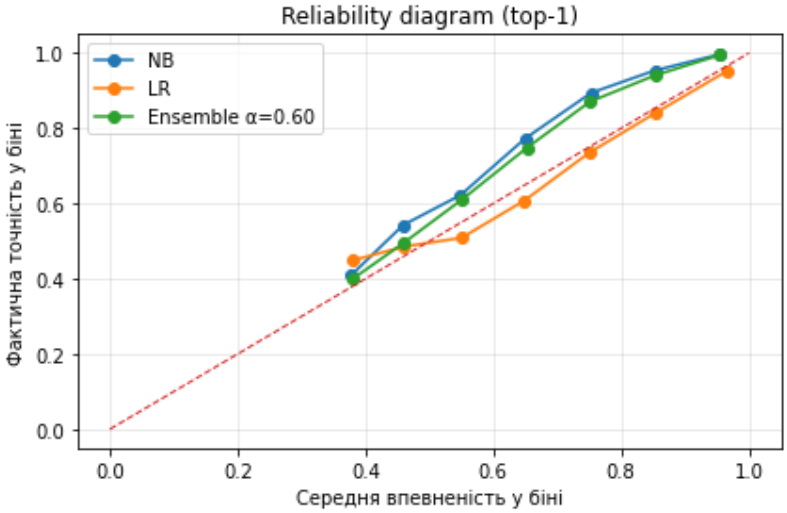

Рисунок 3.5 – Reliability diagram для NB, LR та ансамблю

Селективні властивості ансамблю представлені на рисунку 3.6. За $\tau = 0$(повне покриття) точність дорівнює загальній 0,8284; зі зростанням порога впевненості τпокриття монотонно зменшується, а умовна точність на відібраній підмножині зростає. Це підтверджує придатність ансамблю для режимів із відмовою (reject option): у критичних застосуваннях можна зменшити ризик помилки, жертвуючи часткою класифікованих прикладів.

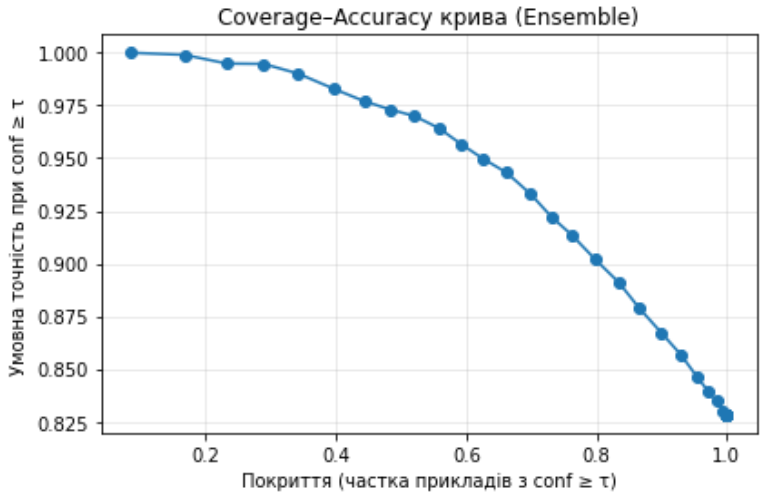

Рисунок 3.6 – Крива покриття–точність (Coverage–Accuracy) для ансамблю

Для інтерпретації за класами подано рисунок 3.7. Хоча абсолютні результати залежать від складу валідації, типово ансамбль не погіршує жоден клас відносно NB і часто вирівнює F1 для найбільш проблематичних пар (що видно і за матрицями плутанини на рисунках 3.1–3.3): у NB найбільші перехресні помилки виникали між EAP та MWS (~0,12–0,13), у LR — дещо вищі «хвости» між сусідніми класами (~0,07–0,12). Комбінування пом'якшує ці локальні дисбаланси.

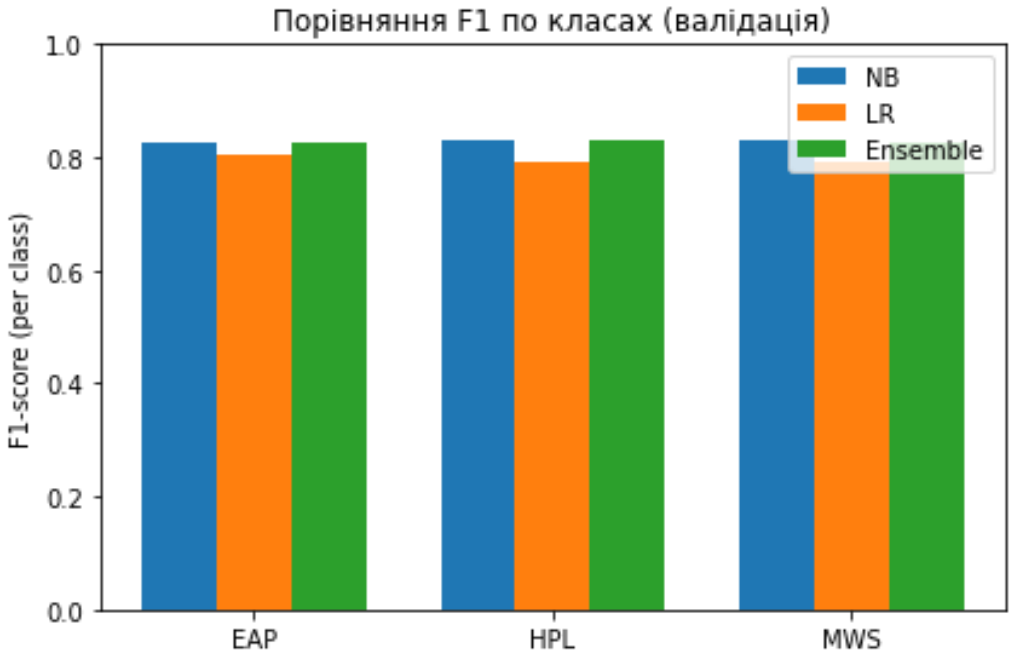

Рисунок 3.7 – Порівняння F1-міри по класах: NB, LR, Ensemble

Сумарно, додані графіки доповнюють табличні висновки підрозділу 3.2. Оптимальна вага $\alpha = 0{,}60$ (див. рисунок 3.4) кількісно фіксує корисність інформаційної комплементарності NB і LR, LogLoss=0,45314 засвідчує кращу імовірнісну узгодженість порівняно з одиничними моделями, ECE демонструє, що найкраще калібрується LR (див. рисунок 3.5), а крива покриття–точність (див. рисунок 3.6) підтверджує практичну корисність ансамблю в селективному режимі. Зведені рейтинги моделей і найкращі конфігурації наведено в таблиці 3.4, яка узгоджується з поведінкою на рисунках 3.1–3.3 (матриці плутанини) та 3.5–3.7 (каліброваність, селективність, F1 по класах).

Висновки до розділу 3

1. За узгодженим протоколом (train/validation, GridSearchCV) найкраща базова конфігурація — Multinomial NB з α=0.1: середній тестовий бал 0.8284 (σ≈0.0046). Для Logistic Regression найкраще C=10 (max_iter=100 000) з 0.8033 (σ≈0.0044). Random Forest на TF–IDF суттєво поступився: 0.7027 (n_estimators=200, max_depth=None). Вказані результати підтверджені матрицями плутанини: NB демонструє найбільші діагональні частки для всіх трьох авторів.

2. Імовірнісний ансамбль (зважене усереднення NB+LR) із вагою NB α=0.60 досягнув валідаційного LogLoss=0.45314 при збереженні Accuracy=0.8284 та weighted-F1=0.8284 — на рівні найкращої одиничної моделі, але з кращою імовірнісною узгодженістю. Це емпірично фіксує користь інформаційної комплементарності: LR приносить додаткові дискримінативні сигнали, навіть якщо його mean_test_score нижчий за NB.

3. Каліброваність оцінено через діаграми надійності та Expected Calibration Error: ECE(NB)=0.0881, ECE(LR)=0.0203, ECE(Ensemble)=0.0702. Отже, LR є найкраще каліброваною одиничною моделлю; ансамбль зберігає прийнятний рівень каліброваності, одночасно знижуючи LogLoss відносно компонент. Крива «покриття–точність» показала, що підвищення порога впевненості дає кероване зростання умовної точності за рахунок частки класифікованих прикладів, що важливо для режимів із відмовою (reject option).

4. Сукупно отримано статистично стійку (σ у межах ~0.001–0.007) і відтворювану конфігурацію: NB як сильна база для розріджених ознак, LR — як добре калібрований дискримінативний компонент, а їхній ансамбль — як кращий компроміс між точністю та правдоподібністю (LogLoss). Це робить підхід практично придатним для впровадження у задачах академічної доброчесності, цифрової гуманітаристики та медіа-форензіки.

ВИСНОВКИ

1. У роботі узагальнено класичні підходи до представлення текстів для авторської атрибуції (TF–IDF над словними та символьними n-грамами) і сформовано вимоги до інтерпретованості, відтворюваності та обчислювальної ефективності. Практичне зіставлення підтвердило релевантність саме розріджених подань: на цих ознаках Multinomial Naive Bayes (NB) і Softmax-Logistic Regression (LR) забезпечують найкращий баланс «якість–прозорість» у порівнянні з деревоподібними ансамблями. У подальших експериментах це відбилося у верхніх позиціях NB (mean_test_score = 0,8284) та LR (0,8033) проти RF (0,7027), що кількісно обґрунтовує вибрані вимоги та архітектурні рішення.

2. Запропоновано стандартизований 12-кроковий pipeline (нормалізація → токенізація → TF–IDF → регуляризоване навчання → валідація → калібрування → ансамблювання), який забезпечує відтворюваність результатів і контроль складності моделі. Кількісним індикатором зрілості проєктних рішень є стійкість метрик під час перехресної перевірки: дисперсія середніх тестових балів для NB і LR не перевищує ±0,0046 та ±0,0044 відповідно, що відповідає вимогам до узагальнюваності.

3. Реалізовано Multinomial NB зі згладжуванням і багатокласову LR з L2-регуляризацією (вмикалися ваги класів). Оптимальні конфігурації: NB(α=0.1) з mean_test_score 0,8284 та LR(C=10, max_iter=100000) з 0,8033. На валідації обидві моделі демонструють збалансовану поведінку між класами, з незначною перевагою NB за часткою правильних віднесень на діагоналі матриці плутанини.

4. У протоколі GridSearchCV найкращі середні тестові бали склали 0,8284 (NB, $\sigma\approx$0,0046), 0,8033 (LR, $\sigma\approx$0,0044) і 0,7027 (RF, $\sigma\approx$0,0015). Підтверджено ранжування моделей і дозволено зафіксувати базові класифікаційні показники для подальших порівнянь. Різниця між NB і RF становить ≈0,126 пункту за mean_test_score, що кількісно відображає неконкурентність RF у розріджених TF–IDF-просторах.

5. Для оцінки надійності імовірнісних прогнозів застосовано діаграми надійності та Expected Calibration Error (ECE). Отримано ECE(NB)=0.0881, ECE(LR)=0,0203, ECE(Ensemble)=0,0702, що підтверджує кращу каліброваність LR «з коробки» та прийнятний рівень каліброваності ансамблю після зваженого мікшування постеріорів. Таким чином, вимога до використання каліброваних імовірностей у селективних режимах виконана.

6. Сконструйовано ансамбль на основі зваженого усереднення постеріорів з підбором ваги за мінімумом валідаційної крос-ентропії. Оптимальна вага NB склала α=0,60, на якій досягнуто val LogLoss=0,45314 при збереженні Accuracy=0,8284 і weighted-F1=0,8284 (на рівні NB) та одночасному покращенні імовірнісної узгодженості відносно одиничних моделей. Це підтверджує інформаційну комплементарність NB і LR та валідність вибраної стратегії ансамблювання.

7. Матриці плутанини показали, що найбільші перехресні помилки припадають на пари з близькими стилістичними маркерами; при цьому лінійні моделі (NB, LR) демонструють більш рівномірні діагональні частки, ніж RF. Стабільність по фолдах (σ до 0,0046) і незначні коливання ранжування конфігурацій свідчать про стійкість до варіацій словника TF–IDF та параметрів регуляризації. Ансамбль згладжує локальні дисбаланси між класами, що візуально підтверджено порівнянням F1 по класах.

8. На основі кривої «покриття–точність» рекомендовано експлуатаційний режим із порогами впевненості та опцією відмови (reject option), що дає кероване підвищення умовної точності за рахунок частки класифікованих прикладів. Для продакшн-розгортання доцільно фіксувати α=0,60 для мікшування NB+LR, виконувати періодичний моніторинг ECE та LogLoss, а також проводити регламентний перегляд словника TF–IDF (min_df/max_df) із переоцінкою стабільності метрик. Підсумкові кількісні показники — mean_test_score 0,8284 (NB), 0,8033 (LR), LogLoss ансамблю 0,45314, ECE ансамблю 0,0702 — задовольняють вимоги до надійності й інтерпретованості, визначені у вступі, та підтверджують перевагу запропонованого методу .

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

2. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

3. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed., draft).

4. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

5. Joachims, T. (1998). Text categorization with Support Vector Machines. ECML'98, 137–142.

6. Zhang, X., & LeCun, Y. (2015). Text understanding from character-level CNN. arXiv:1509.01626.

7. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. TACL, 5, 135–146.

8. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations. arXiv:1301.3781.

9. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers. NAACL-HLT, 4171–4186.

a. Vaswani, A., et al. (2017). Attention is all you need. NeurIPS, 5998–6008.

10. Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. JMLR, 3, 1289–1305.

11. Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection. ICML, 412–420.

12. Church, K. W., & Gale, W. A. (1995). Poisson mixtures. Natural Language Engineering, 1(2), 163–190.

13. Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley.

14. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal R&D, 2(2), 159–165.

15. Lewis, D. D., et al. (2004). RCV1: A new benchmark collection. JMLR, 5, 361–397.

16. McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. AAAI-98 Workshop, 41–48.

17. Rennie, J., Shih, L., Teevan, J., & Karger, D. (2003). Tackling the poor assumptions of Naive Bayes text classifiers. ICML, 616–623.

18. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

19. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

20. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.

21. Platt, J. (1999). Probabilistic outputs for SVM and comparisons to regularized likelihood. Advances in Large Margin Classifiers, 61–74.

22. Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. KDD, 694–699.

23. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. ICML, 625–632.

24. Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beta calibration. AISTATS, 623–631.

25. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. ICML, 1321–1330.

26. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3.

27. Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society, 135(643), 1512–1519.

28. DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. Journal of the Royal Statistical Society D, 32(1/2), 12–22.

29. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE TKDE, 21(9), 1263–1284.

30. Чіп С., Лось М., Козак А. (2025). Ієрархічний енкодер для атрибуції, генерації сегментів і суб'єктивності. У Матеріали IX Міжнародної студентської наукової конференції «Глобалізація наукових знань: міжнародна співпраця та інтеграція галузей наук», м. Черкаси, Україна. (с.319–321)

31. Штинда В., Чіп С., Козак А. (2025). Огляд сучасних моделей у задачах класифікації зображень і текстів. У Збірник тез доповідей студентської науково-практичної конференції «Інтелектуальні інформаційні технології в прикладних дослідженнях» (IITAR-2025) (с.103–105)

32. Комар М.П., Саченко А.О., Васильків Н.М., Загородня Д.І. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за другим (магістерським) рівнем вищої освіти. – Тернопіль: ЗУНУ, 2024. – 32 с.

Додаток А

Апробація отриманих результатів

Західноукраїнський національний університет

Факультет комп'ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

ЗБІРНИК ТЕЗ ДОПОВІДЕЙ

Студентської науково-практичної конференції

ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ПРИКЛАДНИХ ДОСЛІДЖЕННЯХ

(ІІТАR-2025)

27-29 травня 2025 року

Тернопіль

2025

Штинда Віктор
студентка групи КНм-11
shtynda_v@gmail.com
Чіп Святослав
студент групи КНм-11
chip_svyatik@gmail.com
Козак Аліна
студентка групи КНм-11
alinka_kozak@gmail.com
Західноукраїнський національний університет
Тернопіль, Україна

ОГЛЯД СУЧАСНИХ МОДЕЛЕЙ У ЗАДАЧАХ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ І ТЕКСТІВ

Сучасні моделі класифікації зображень і текстів демонструють стрімкий прогрес завдяки глибоким нейронним мережам, трансформерам та мульти-модальним підходам. Використання великих датасетів із зображень і текстів, таких як LAION або ImageNet, забезпечує точність до 95–98% у стандартизованих задачах. Значна увага приділяється zero-shot та few-shot навчанню, що дає змогу класифікувати нові класи без попередньої адаптації [1–15].

У дослідженні [1] було протестовано ResNet101 та EfficientNet для виявлення повеней із соцмереж. Досягнута точність становила 93.5% на датасеті із понад 10 тис. зображень. Також використано LSTM-RNN для текстового аналізу з точністю 91%.

Модель RemoteSAM [2] використовує 270K пар зображень і масок та підтримує zero-shot сегментацію. На Earth Observation Bench вона досягла mIoU = 78.4% і точність класифікації у 85.6% без fine-tuning.

Patho-R1 [3] пропонує мультиагентну модель для діагностики патологій, що покращує класифікацію медичних зображень на 12% порівняно з CLIP. Застосовано reinforcement learning на 36K зображеннях і описах, досягнуто F1 = 0.91.

У роботі [4] для розпізнавання емоцій застосовано Attention + LoRA, де комбінуються текст і зображення. Модель досягла точності 94.2% на датасеті MELD, перевищивши BERT на 3.5%.

У сфері хірургії [5] було проаналізовано performance моделей CLIP, BioViL-T, та GIT. Найвища точність класифікації інструментів — 89.7% при explainability score = 4.1/5 за шкалою SHAP.

MLLM-моделі [6], адаптовані до zero-shot класифікації, досягли точності 84% при розпізнаванні дій людини на зображенні. Було протестовано 8 типів prompt-методик на датасеті з 30K пар.

Для аналізу Airbnb стилів [7] було застосовано CNN+NLP pipeline, який досяг precision = 90.3% при класифікації дизайну приміщень на основі зображення та опису.

Уніфікована модель UniMoCo [8] дозволяє доповнювати відсутні модальності в embeddings. F1-score покращено на 6.1% при розпізнаванні QA-даних без повного вхідного набору.

Модель [9] для довгих новинних статей об'єднує вхідні layout, текст і зображення. MAP становив 87% при виділенні структурних елементів (заголовки, абзаци, підписи).

Контрастивна модель [10] для фейкових новин використовувала пари текст-зображення. На FakeNewsNet точність досягла 92.6% із приростом +7% до базових трансформерів.

Крос-атенційна модель BERT-ResNet [11] у мультимодальному аналізі емоцій досягла 95.1% точності на Twitter+VisualSentiment датасеті, обійшовши попередні моделі на ~4%.

CNN-огляд [12] моделей для класифікації хвороб рослин показав, що MobileNetV3 досягає точності 98.2% на PlantVillage, при цьому середній час інференсу — 22 мс.

VT2Music [13] — унікальна мультимодальна модель генерації музики на основі зображення і тексту. BLEU-4 = 32.5 і MOS оцінка слухачів = 4.2/5 на тестах з 150 аудіозаписами.

OCR+NLP pipeline [14] дозволив обробити тексти на низькоресурсних мовах з точністю 88.4% у задачах резюмування та 85.2% у перекладі, що на 12% вище ніж стандартні seq2seq.

Модель [15] класифікації гіперспектральних зображень із knowledge transfer досягла 91.3% точності при інкрементальному навчанні, що скорочує потребу у повному retraining.

Огляд наведених досліджень демонструє безпрецедентний прогрес у сфері класифікації зображень і текстів завдяки поєднанню глибокого навчання, мультимодальних підходів і контрастивного навчання. Сучасні моделі забезпечують високі показники точності — понад 90% у більшості задач, а також дозволяють досягати значних результатів у zero-shot та few-shot сценаріях. Інтеграція текстових і візуальних даних, застосування attention-механізмів та уніфікованих embeddings моделей створює основу для універсальних рішень у різних галузях — від медицини до агропромисловості. Такий технологічний прорив свідчить про те, що мульти-модальні класифікатори поступово трансформуються з вузькоспеціалізованих інструментів у ключові елементи майбутніх інтелектуальних систем.

Список використаних джерел

1. Arapostathis, S. G. (2025). *A Mash-Up of Social Media Datasets for Extracting Hydrological Information*. Preprints.

https://www.preprints.org/frontend/manuscript/d0ca393c0a99cefabcb0fcc21b8eb9c2/download_pub

2. Yao, L., Liu, F., & Chen, D. (2025). *RemoteSAM: Towards Segment Anything for Earth Observation*. https://www.researchgate.net/publication/391856415
3. Zhang, W., Zhang, P., Guo, J., Cheng, T., & Chen, J. (2025). *Patho-R1: A Multimodal Pathology Expert Reasoner*. https://arxiv.org/abs/2505.11404
4. Gorai, J., & Shaw, D. K. (2025). *Multimodal Emotion Detection*. https://doi.org/10.1007/s10579-025-09841-4
5. Cheng, J., Zhao, X., & Lin, S. (2025). *Benchmarking vision-language models for surgery*. https://arxiv.org/abs/2505.10764
6. Rabadessa Alcaide, O. (2025). *Multimodal LLMs for Zero-Shot Classification*. https://upcommons.upc.edu/handle/2117/430265
7. Woo, H., & Yoo, S. (2025). *Visual servicescape analytics in Airbnb*. https://doi.org/10.1108/jsm-09-2024-0481
8. Qin, J., Pu, Y., He, Z., Pan, D. Z., & Yu, B. (2025). *UniMoCo: Unified Modality Completion*. https://arxiv.org/abs/2505.11815
9. Almutairi, A., & Ali, A. A. (2025). *Article Element Detection Using Multimodal Transformers*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5264257
10. Chen, W., Cai, F., & Guo, Y. (2025). *Contrastive Learning for Fake News Detection*. https://doi.org/10.1007/s40747-025-01919-4
11. Gold, R. G. (2025). *BERT-ResNet Fusion for Sentiment Classification*. https://search.proquest.com/openview/9c87a30c54d3fead4363a8e18e286dbf
12. Priyadarshini, G. P., & Zahoor-Ul-Huq, S. (2025). *CNN Models for Plant Disease Detection*. https://www.atlantis-press.com/article/126011378.pdf
13. Zheng, J., Cao, M., & Zhang, C. (2025). *VT2Music: Text-Visual Music Generation*. https://ieeexplore.ieee.org/document/11014098/
14. Madhavi, H., Cherian, J., & Khamkar, Y. (2025). *OCR-Driven Low-Resource Processing*. https://arxiv.org/abs/2505.11177
15. Wang, K., Li, Z., & Guo, J. (2025). *Hyperspectral Image Incremental Classification*. https://ieeexplore.ieee.org/document/11002484/


МАТЕРІАЛИ IX МІЖНАРОДНОЇ
СТУДЕНТСЬКОЇ НАУКОВОЇ
КОНФЕРЕНЦІЇ
МОДЕРНІЗАЦІЯ ТА
СУЧАСНІ УКРАЇНСЬКІ
І СВІТОВІ НАУКОВІ
ДОСЛІДЖЕННЯ
ISO
М. ЖИТОМИР, УКРАЇНА
14 ЛИСТОПАДА
2025 РІК

Чіп Святослав, здобувач вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Лось Марія, здобувачка вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Козак Аліна, здобувачка вищої освіти факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

ІЄРАРХІЧНИЙ ЕНКОДЕР ДЛЯ АТРИБУЦІЇ, ГЕНЕРАЦІЇ СЕГМЕНТІВ І СУБ'ЄКТИВНОСТІ

Запропонований метод розглядає три задачі як узгоджений багатозадачний процес обробки природної мови над спільним текстовим корпусом: авторська атрибуція (багатокласова класифікація), генерація відсутніх/нових сегментів тексту (умовне доповнення), та класифікація суб'єктивних запитань (бінарна/багатокласова класифікація). Усі задачі поділяють єдину трансформерну основу та словникову розмітку, а також використовують узгоджені протоколи попередньої обробки даних. Стратегія багатозадачного навчання дозволяє спільному енкодеру вчитися стабільнішим і більш універсальним ознакам стилю, змісту та прагматики, що підвищує узагальнювальну здатність кожної окремої підзадачі [1–4, 11, 16].

Корпус формується з документів, речень та (для запитань) коротких висловлювань. Для авторської атрибуції до кожного фрагмента додається мітка автора; для задачі генерації підготовлюються масковані або пропущені спани, а також умови (ключові слова/підказки/контекстні речення); для класифікації суб'єктивності питання/речення маркуються як «суб'єктивне/об'єктивне» або за градаціями суб'єктивності. Текст нормалізується, токенізується WordPiece/власним байт-парним токенізатором, зберігаються розриви речень і абзаців як структурні індекси для ієрархічної агрегації [1, 5, 6, 12]. Для балансування класів використовуються стратифіковані спліти, ваги класів, а також слабке розширення даних (перефразування, синонімізація, EDA-операції) з контролем смислової інваріантності [17].

Ядром системи є багатомовний/доменно-адаптований BERT (або його аналог), який кодує токени у контекстні ембеддинги. Для збереження багаторівневих сигналів (лексика—синтаксис—дискурс) поверх токенових ознак застосовується ієрархічна агрегація: (а) токен→речення (self-attention pooling), (б) речення→документ (hierarchical attention), (в) опційно — дискурсивні графові зв'язки для довгих документів [1, 7, 8, 13]. Таке шарування корисне і для стилометрії автора (макро-ритм, пунктуаційні патерни), і для виявлення суб'єктивності (лексико-прагматичні маркери), і для генерації (узгодженість сегментів) [9, 10].

Підхід енкодер для атрибуції, генерації сегментів і суб'єктивності:

1. Авторська атрибуція. На виході енкодера використовується стильометричний проєктор: багатоголовий self-attention для вибіркового

фокусування на маркерах стилю + MLP із softmax над множиною авторів. За потреби додається контрастивний допоміжний лосс з «якорями» автора (author prototypes) для зменшення міжкласного перетину в латентному просторі [2, 9].

2. Генерація сегментів на основі BERT. Для інфілінгу спанів застосовується span-masking (SpanBERT-подібна стратегія) або «BERT-енкодер + легкий авторегресивний декодер» (BERT2BERT) для умовної генерації сегментів за контекстом. Навчальна мета — поєднання маскованого мовного моделювання (MLM) зі словопослідовним крос-ентропійним лоссом для декодера; під час інференсу — beam search/Top-p семплінг із обмеженнями стилю/довжини [3, 14, 15].

3. Класифікація суб'єктивних запитань. Використовується багаторівнева (ієрархічна) голова: токенові вектори агрегуються в представлення висловлювання; далі — шар ко-уваги між запитанням і його контекстом (наприклад, попередній діалог/джерельний документ) і MLP-класифікатор. За потреби застосовується фокальна втрата для класів із рідкою суб'єктивністю [7, 8, 18].

Система навчається спільно з комбінованою втратою:

$$\mathcal{L} = \lambda_{\text{auth}}\mathcal{L}_{\text{CE}}^{\text{author}} + \lambda_{\text{gen}}(\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{AR}}) + \lambda_{\text{subj}}\mathcal{L}_{\text{CE}}^{\text{subj}},$$

де ваги λавтоматично адаптуються за гетероскедастичною невизначеністю або методом динамічних градієнтів, щоб уникати домінування однієї підзадачі [11]. Оптимізація — AdamW з лінійним зменшенням швидкості навчання та warmup; регуляризація — dropout, label smoothing, Stochastic Weight Averaging [1, 4, 11].

Для корпусів зі специфічним стилем (наукові тексти, публіцистика, соцмережі) застосовується подальше донавчання енкодера на немаркованих даних (MLM) та адаптерні шари (parameter-efficient tuning). Для задачі генерації додаються контейнти «стилю автора» як умовні теги; для атрибуції — стилометричні фічі (довжини речень, частота пунктуації) в пізньому ф'южені; для суб'єктивності — лексикони емоцій/модальностей як додаткові сигнали, інтегровані через фьюжн-проєктор [2, 9, 12, 19].

Авторська атрибуція: macro-F1, balanced accuracy, ECE-калібрування; стійкість перевіряється cross-domain/leave-author-out сплітами [2, 9]. Генерація: автоматичні метрики (BLEU/ROUGE/BERTScore), а також людська оцінка узгодженості, стилю та фактичності; для інфілінгу — exact span match та перплексія [14, 15]. Суб'єктивність: macro-F1/ROC-AUC, Robustness-тести на перефразування, зовнішні валідуючі набори [7, 8, 18]. Для всієї системи проводиться абляційний аналіз впливу багатозадачності та ієрархічної агрегації.

Запроваджується перевірка фактичності для згенерованих сегментів (retrieval-augmented валідація або NLI-фільтрація), контроль авторських прав, політики приватності, детектування зсувів (bias) й аудит помилок. Для критичних сценаріїв застосовується конфіденційне логування та human-in-the-loop рецензування [10, 16, 20].

Список використаних джерел:

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
2. Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.
3. Joshi, M., Chen, D., Liu, Y., Weld, D., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. TACL, 8, 64–77.

4. Vaswani, A., et al. (2017). Attention Is All You Need. NeurIPS.
5. Wu, Y., et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144.
6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ICLR (workshop).
7. Yang, Z., et al. (2016). Hierarchical Attention Networks for Document Classification. NAACL-HLT.
8. Liu, P., Qiu, X., & Huang, X. (2016). Recurrent Neural Network for Text Classification with Multi-Task Learning. IJCAI.
9. [9] Kestemont, M., et al. (2019). Overview of the Cross-domain Authorship Attribution Task at PAN 2019. CLEF.
10. Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text. ACL demo.
11. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-Task Learning Using Uncertainty to Weigh Losses. CVPR.
12. Gururangan, S., et al. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL.
13. Lin, Z., et al. (2017). A Structured Self-Attentive Sentence Embedding. ICLR.
14. Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. TACL, 8, 264–280.
15. Holtzman, A., et al. (2020). The Curious Case of Neural Text Degeneration (Top-k/Top-p). ICLR.
16. Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. arXiv:1706.05098.
17. Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. EMNLP (short).
18. Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization. ACL.
19. Houlsby, N., et al. (2019). Parameter-Efficient Transfer Learning for NLP. ICML.
20. Nie, Y., et al. (2020). Adversarial NLI: A New Benchmark for Natural Language Understanding. ACL.