УДК

С.В. ЗАБОЛОТНІЙ
Черкаський державний бізнес коледж, Черкаси, Україна
e-mail: zabolotniua@gmail.com
А.В. ЧЕПИНОГА
Черкаський державний технологічний університет, Черкаси, Україна
e-mail: a.chepynoha@chdtu.edu.ua
В.І. ХОТУНОВ
Черкаський державний бізнес коледж, Черкаси, Україна
e-mail:

ВІД СТАТИСТИЧНОГО РОЗПІЗНАВАННЯ ОБРАЗІВ ДО АНАЛІЗУ ЕМОЦІЙ: ЗАСТОСУВАННЯ АПАРАТУ РОЗКЛАДУ В ПРОСТОРІ ІЗ ПОРІДНИМ ЕЛЕМЕНТОМ ДО МОДЕЛЕЙ ОБРОБКИ ПРИРОДНОЇ МОВИ

Анотація. Розпізнавання емоцій у текстах є важливою задачею сучасної обробки природної мови, де на даний момент домінують трансформерні архітектури. Однак їхні внутрішні механізми залишаються «чорною скринькою», а якість класифікації, особливо для складних випадків, має потенціал для покращення. У цій роботі пропонується новий гібридний підхід, що поєднує потужність сучасних мовних моделей з глибоким аналізом їхніх векторних представлень шляхом адаптації класичного методу статистичного розпізнавання образів, заснованого на розкладі в просторі з порідним елементом (просторі Кунченка). Метод генерує новий набір «статистико-геометричних» ознак, що базуються на похибці реконструкції векторного текстових повідомлень відповідних класів.
Експерименти на українському (EMOBENCH-UA) та англійському (EmoEvent) наборах даних показали, що запропонований гібридний підхід статистично значуще покращує якість класифікації. Дослідження також виявило ключові умови ефективності методу: він є потужним «уточнювачем» для моделей, донавчених на цільовій задачі, але неефективний на «сирих», неспеціалізованих векторних представленнях. Встановлено, що вибір базисних функцій для реконструкції є важливим гіперпараметром, що дозволяє адаптувати метод до специфічної геометрії простору даних.

Ключові слова: розпізнавання емоцій, обробка природної мови, векторне представлення, простір Кунченка, генерація ознак, гібридна модель.

ВСТУП

Автоматичне розпізнавання емоцій у текстах (Emotion Detection) є однією з найбільш актуальних та водночас складних задач обробки природної мови (NLP) [1]. Можливість точно ідентифікувати емоційний стан людини відкриває широкі перспективи у таких сферах, як аналіз відгуків клієнтів, моніторинг психічного здоров'я, політичні науки, а також створення адаптивних людино-машинних інтерфейсів.

Сучасний етап розвитку NLP характеризується домінуванням глибоких нейронних мереж, зокрема трансформерних архітектур, таких як BERT (Bidirectional Encoder Representations from Transformers) [2] та його численні варіації, наприклад, RoBERTa [3] та XLM-RoBERTa [4]. Ці моделі продемонстрували передові результати в широкому спектрі задач завдяки своїй здатності вловлювати складні контекстуальні залежності в тексті та формувати на їх основі щільні багатовимірні векторні представлення — ембедінги. Саме в цих векторних просторах, як припускається, і закодована вся необхідна семантична інформація для вирішення цільової задачі.

Однак, попри високу ефективність, трансформерні моделі мають і свої обмеження. По-перше, вони часто сприймаються як «чорні скриньки», а внутрішня геометрія та структура їх векторних просторів залишаються предметом активних досліджень [5, 6].

По-друге, навіть найкращі моделі не завжди здатні розрізняти тонкі нюанси між близькими за семантикою емоціями (наприклад, гнів чи огида), що залишає простір для подальших покращень. Для вирішення цих проблем часто застосовуються гібридні підходи, що поєднують потужність глибоких моделей з іншими джерелами інформації. Зазвичай такі підходи доповнюють нейромережеві ознаки даними з емоційних лексиконів або іншими лінгвістичними ознаками [7]. Існуючі методи аналізу векторних представлень, такі як діагностичні класифікатори (probing classifiers) [8], зосереджені на виявленні того, яка лінгвістична інформація закодована у відповідному векторному просторі, а не на покращенні основної задачі [9]. З іншого боку, методи, що базуються на зниженні розмірності (наприклад, t-SNE), використовуються частіше для візуалізації та не створюють ознаки, оптимізовані для класифікації. Таким чином, потенціал використання ознак, що описують класо-специфічну геометрію простору векторного представлення текстових патернів, залишається менш дослідженим.

Це спонукало звернутися до класичного підходу, заснованого на використанні методів статистичного розпізнавання образів, що були розроблені для аналізу багатовимірних даних. Такі фундаментальні математичні апарати, як розклади в ряди за ортогональними базисними функціями (ряди Фур'є, поліноми Ерміта та ін.) або методи зниження розмірності (PCA, розклад Карунена-Лоєва), десятиліттями слугували основою для вирішення різноманітних практичних задач [10]. Серед таких математичних апаратів особливе місце посідає метод розкладу в просторі з порідним елементом [11], розроблений Ю.П. Кунченко в рамках його теорії стохастичних поліномів [12]. Цей розклад характеризується ключовою особливістю, яка полягає у відсутності вимоги ортогональності до системи нелінійних базисних функцій. Результати ряду проведених досліджень підтвердили потенційну ефективність застосування цього апарату при вирішенні задач статистичного опрацювання даних у випадках негаусовості їх розподілу. Зокрема, в роботі [13] був теоретично обґрунтований новий метод статистичного розпізнавання образів та було продемонстровано його здатність зменшувати дисперсію ознак і покращувати точність класифікації. У роботі [14] наведена методологія використання розкладу Кунченко при вирішенні завдань ймовірнісної діагностики розладки негаусових випадкових процесів. Дослідження [15] продемонструвало перспективність для перевірки статистичних гіпотез про середні значення, показавши значне зменшення ймовірності помилок другого роду.

У даній роботі ми висуваємо гіпотезу, що цей, здавалося б, суто статистичний метод може бути успішно адаптований для аналізу векторного представлення текстових повідомлень з метою покращення якості розпізнавання емоцій. Ми прагнемо збудувати міст між двома парадигмами: сучасною, що базується на навчанні глибоких нейромережевих моделей, та класичною, що фокусується на ретельному конструюванні інформативних ознак.

Основні питання, на які ми шукаємо відповідь:

- Чи може гібридна модель, що поєднує виходи трансформера та ознаки, згенеровані на основі похибки реконструкції в просторі Кунченка, перевершити базову трансформерну модель?
- Наскільки інформативними є згенеровані «статистико-геометричні» ознаки самі по собі?
- Які умови є необхідними для ефективного застосування запропонованого методу?
- Як вибір базисних функцій для реконструкції впливає на кінцеву якість класифікації?

Для відповіді на ці питання ми проводимо серію експериментів на двох різномовних наборах даних: українському EMOBENCH-UA [16] та англійському EmoEvent [17],

порівнюючи запропонований підхід з низкою еталонних тестів.

1. ТЕОРЕТИЧНІ ОСНОВИ МЕТОДУ

Запропонований у цьому дослідженні підхід до генерації класифікаційних ознак ґрунтується на ключовій інновації розкладу Кунченко, яка полягає у формуванні різниці між порідним елементом (вектором у багатовимірному просторі) та його апроксимаційним представленням у вигляді полінома у формі лінійної комбінації суми вагових коефіцієнтів, помножених на нелінійні базисні перетворення від цього ж порідного елемента. Критерієм оптимізації вагових коефіцієнтів виступає мінімум середньоквадратичної похибки (СКП) розкладу (різниці між порідним елементом та його апроксимацією). Таким чином, цей апарат дозволяє апроксимувати складні, нелінійні залежності у багатовимірних даних за допомогою стохастичних (у сенсі усереднення результатів функціональних перетворень) поліномів, що робить його перспективним інструментом для аналізу щільних векторних представлень (ембедингів), які генерують сучасні мовні моделі.

1.1. Поліноміальне наближення у просторі з порідним елементом

Нехай векторне представлення (ембединг) тексту представлено як D-вимірний випадковий вектор $X = \{x_1, x_2 ..., x_D\}$. Основна ідея методу полягає в тому, щоб наблизити цей «порідний» вектор X іншим вектором Y такої ж розмірності, компоненти якого формуються як узагальнені поліноми S-го порядку від компонент вектору X:

$$y_n = k_0 + \sum_{i=1}^{S} k_i \cdot \phi_i(x_n), \ для\, n = 1, ..., D, \tag{1}$$

де:

k_0 та k_i — скалярні коефіцієнти полінома, що є сталими для всіх компонент вектору.

$\phi_i(\cdot)$ — набір з S наперед визначених, загалом кажучи, нелінійних базисних функцій.

x_n та y_n — n-ті компоненти векторів X та Y відповідно.

Важливо зазначити, що реконструкція кожної компоненти y_n залежить лише від відповідної компоненти x_n. Це означає, що модель реконструкції є **покомпонентною**, а коефіцієнти k_0 та k_i є скалярними величинами, спільними для всіх D компонент.

Для ілюстрації, процес можна представити у векторній формі. Якщо застосувати оператор $\Phi(\cdot)$ до вектору X покомпонентно, щоб отримати матрицю $\Phi(X)$ розмірності $S \times D$, то вираз (1) можна умовно записати як:

$$Y = k_0 1^T + K^T \Phi(X), \tag{2}$$

де

$K = [k_1, k_2, ..., k_S]^T$ — вектор-стовпець коефіцієнтів, а 1 — вектор з одиниць розмірності D.

Таким чином єдиний набір коефіцієнтів K застосовується до всього простору ембедингів.

Вибір набору базисних функцій $\phi_i(\cdot)$є ключовим аспектом методу. На відміну від класичних підходів, що часто спираються на ортогональні системи (ряди Фур'є, поліноми Ерміта, Лежандра), апарат простору з порідним елементом не накладає вимоги ортогональності. Це дозволяє використовувати значно ширший клас функцій, зокрема неортогональні степеневі, дробно-степеневі, тригонометричні чи інші нелінійні

перетворення, які можуть краще відповідати специфічній геометрії конкретного простору даних.

1.2. Критерій оптимальності та знаходження коефіцієнтів розкладу

Оптимальні коефіцієнти розкладу k_0 та K знаходяться з умови мінімізації СКП між порідним вектором X та його апроксимацією Y:

$$E[\| X - Y \|^2] \to min \tag{3}$$

де $E[\cdot]$ — оператор математичного сподівання. Як показано в [11], для забезпечення мінімуму похибки, вектор коефіцієнтів K та коефіцієнт зміщення k_0 мають задовольняти наступним умовам.

Спочатку, для конкретного класу m, на навчальній вибірці, що містить лише приклади цього класу, обчислюється вектор коефіцієнтів $K^{(m)}$, що є розв'язком системи лінійних алгебраїчних рівнянь:

$$F^{(m)} \cdot K^{(m)} = B^{(m)}, \tag{4}$$

де:

$F^{(m)}$ — коваріаційна матриця базисних функцій розмірності $S \times S$ з елементами $F_{ij} = E[(\phi_i(x) - E[\phi_i(x) \mid m]) \cdot (\phi_j(x) - E[\phi_j(x) \mid m]) \mid m]$, обчислена на даних класу m.

$B^{(m)}$ — вектор коваріацій між базисними функціями та самою випадковою величиною з елементами $B_i = E[(x - E[x \mid m]) \cdot (\phi_i(x) - E[\phi_i(x) \mid m]) \mid m]$, обчислений на даних класу m.

Після знаходження вектору $K^{(m)}$ (наприклад, методом $K^{(m)} = (F^{(m)})^{-1} B^{(m)}$), відповідний коефіцієнт зміщення $k_0{}^{(m)}$ обчислюється так, щоб забезпечити незміщеність оцінки для даного класу:

$$k_0{}^{(m)} = E[x \mid m] - (K^{(m)})^T \cdot E[\Phi(x) \mid m], \tag{5}$$

де $E[\cdot \mid m]$ — оператор математичного сподівання, обчислений за розподілом даних класу m.

1.3. Формування класифікаційних ознак на основі похибки реконструкції

Ми адаптуємо цей теоретичний апарат для вирішення задачі класифікації текстових патернів, представлених у векторному просторі. Замість того, щоб шукати єдину найкращу апроксимацію для всіх даних, ми висуваємо гіпотезу, що кожен клас емоцій має свою унікальну «геометричну» структуру в просторі ембедингів, а отже, для кожного класу існує своя оптимальна (в рамках обраного базису розкладу) модель реконструкції. Процедура генерації ознак складається з двох етапів:

1. Етап навчання: Для кожного з M класів емоцій $(m = 1, ..., M)$ ми використовуємо відповідну частину навчальної вибірки для обчислення унікального набору параметрів моделі реконструкції $\{K^{(m)}, k_0{}^{(m)}\}$ за формулами (4) та (5). Таким чином, ми отримуємо M класо-специфічних моделей реконструкції $Y^{(m)}$.

2. Етап формування ознак: Для нового тексту, представленого ембедингом X_{new}, ми послідовно застосовуємо до нього кожну з M навчених моделей і обчислюємо середньоквадратичну похибку реконструкції для кожної з них:

$$MSE^{(m)} = (1 / D) \cdot \| X_{new} - Y^{(m)} \|^2. \tag{6}$$

Інтуїтивно, якщо X_{new} дійсно належить до класу m, то похибка $MSE^{(m)}$ буде малою, оскільки модель $Y^{(m)}$ «добре знає», як реконструювати об'єкти свого класу. І навпаки, похибка буде великою при застосуванні «чужої» моделі реконструкції.

Отримані M значень похибок, після логарифмування (для стабілізації дисперсії),

формують новий вектор ознак:

$$F_K(X_{new}) = [log(MSE^{(1)} + \varepsilon), log(MSE^{(2)} + \varepsilon), ..., log(MSE^{(M)} + \varepsilon)] \quad (7)$$

де ε — невелика додатна константа (наприклад, 10^{-9}), що додається для забезпечення обчислювальної стійкості та уникнення логарифмування нуля у випадку ідеальної реконструкції.

Кожна компонента цього вектору F_K несе інформацію про «відстань» або «нетиповість» вхідного тексту відносно центру та структури кожного з класів емоцій. Саме цей вектор і використовується в наших експериментах як набір нових, «геометричних» ознак для подачі на вхід простому класифікатору (наприклад, логістичній регресії) або для об'єднання з вихідними даними трансформера.

2. ДИЗАЙН ЕКСПЕРИМЕНТІВ

Для перевірки запропонованого методу та валідації основних гіпотез дослідження було розроблено дизайн експериментів на прикладі завдання класифікації людських емоцій, що охоплює два різних мовних домени, декілька базових моделей та сценаріїв їх застосування.

2.1. Набори даних

У дослідженні було використано два публічних, добре анотованих набори, розташованих у відкритому доступі на платформі Hugging Face:

- **EMOBENCH-UA** (ukr-detect/ukr-emotions-binary) [16]: Це еталонний набір даних для розпізнавання емоцій в українськомовних текстах. Його було створено шляхом багатоетапної фільтрації твітів з подальшою анотацією на платформі Toloka.ai. Процес анотації було розділено на два окремі проєкти для зменшення когнітивного навантаження на анотаторів, а фінальна якість розмітки характеризується високим показником узгодженості (Krippendorff's alpha = 0.85). Датасет містить 6 основних емоцій та мітку «None» і представлений у багатозначному форматі (multilabel), де один текст може містити кілька емоцій. У нашому дослідженні ми використовували як оригінальну багатозначну постановку, так і спрощену, з однією домінуючою емоцією на текст.
- **EmoEvent** (fmplaza/EmoEvent) [17]: Це багатомовний корпус, з якого для нашого дослідження було взято англомовну частину. Дані складаються з твітів, зібраних навколо резонансних подій у квітні 2019 року (наприклад, пожежа в соборі Нотр-Дам, вибори в Іспанії), та анотовані на платформі Amazon Mechanical Turk. Датасет анотовано за 7 емоціями (6 базових + «Інше») у форматі багатокласової класифікації (multiclass), де кожному тексту присвоєно лише одну емоційну мітку.

Використання цих двох датасетів дозволило перевірити узагальнювальну здатність методу на різних мовах, з різними типами анотації та різними джерелами даних.

2.2. Базові моделі та еталонні тести

Для кожного набору даних було визначено базову модель класифікатора (еталонний тест), з якою порівнювався запропонований гібридний підхід.

1 **Для EMOBENCH-UA:** В якості еталонного тесту використовувалася модель від авторів набору даних ukr-detect/ukr-emotions-classifier [16] — побудована на основі XLM-RoBERTa та спеціально донавчена на українських емоційних даних. Її вихідні ймовірності для кожного класу, оброблені за допомогою оптимально підібраних порогів, слугували точкою відліку для порівняння.

2 **Для EmoEvent:** Експеримент проводився у двох сценаріях для дослідження умов ефективності методу:

Сценарій А (без донавчання): Використовувалася загальноцільова модель

RoBERTa-base без донавчання. Ембединги з цієї моделі слугували входом для нашого методу та для класичних підходів (TF-IDF, PCA). Цей сценарій є зіставним з базовими експериментами, проведеними авторами набору даних, які також використовували класифікатор SVM на лінгвістичних ознаках.

Сценарій Б (з донавчанням): Модель RoBERTa-base була донавчена на навчальній частині датасету EmoEvent. Результати цього донавченого класифікатора стали додатковим сильним еталоном для перевірки, чи може наш метод покращити вже добре налаштовану модель.

2.3. Реалізація методу генерації ознак

Процес генерації ознак Кунченка включав наступні кроки:

1) **Отримання ембедингів:** Для кожного тексту з набору даних ми отримували векторне представлення, використовуючи вихід останнього прихованого шару відповідної базової моделі, що відповідає спеціальному токену [CLS]. Розмірність ембедингів для обох моделей становила D=768.
2) **Вибір базисних функцій**: Спираючись на досвід попередніх досліджень, ми зосередилися на перевірці ефективності використання неортогонального базису із степеневих $\phi_p(x) = x^p$ та дрібно-степеневих перетворень: $\phi_p(x) = sign(x)x^{1/p}$, $p = 2..S$. Вибір величини S є компромісом між представницькою здатністю моделі розкладу та її обчислювальною складністю. З одного боку, більша кількість функцій дозволяє моделювати більш складні залежності, з іншого — призводить до збільшення розмірності коваріаційної матриці F та потенційної обчислювальної нестійкості. Попередні експерименти показали, що $S = 5$ є достатнім для досягнення значущих результатів, хоча формальна оптимізація цього гіперпараметра може стати предметом окремого дослідження.
3) **Навчання та трансформація:** Для кожного набору даних на навчальній вибірці навчалися класо-специфічні моделі реконструкції, які потім застосовувалися до тестової вибірки для генерації фінального вектора з M ознак похибки.

2.4. Метрики та оцінка

- **Основна метрика якості:** Для оцінки всіх моделей використовувався усереднений Macro F1-score, який є стійким до дисбалансу класів і однаково враховує якість розпізнавання кожної емоції.
- **Статистична значущість:** Для порівняння результатів гібридної моделі та бенчмарку застосовувався бутстреп-тест (bootstrap test). Генерувалося 1000 бутстреп-вибірок з тестових даних, на яких обчислювалася різниця в F1-score. На основі розподілу цих різниць розраховувався p-value та 95% довірчий інтервал для покращення. Покращення вважалося статистично значущим при p-value < 0.05.
- **Візуальний аналіз:** Для якісної оцінки простору згенерованих ознак використовувався метод t-SNE, що дозволяв візуалізувати, наскільки добре ознаки різних методів розділяють класи емоцій у двовимірному просторі.

3. РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

У цьому розділі представлено результати експериментального дослідження, проведеного згідно з методологією, описаною в розділі 2. Аналіз результатів організовано навколо ключових дослідницьких питань: ефективність гібридного підходу, інформативність згенерованих ознак та умови їх застосування.

3.1. Ефективність гібридного підходу на спеціалізованих моделях (EMOBENCH-UA)

Перший етап дослідження був спрямований на перевірку основної гіпотези на українському набору даних EMOBENCH-UA, використовуючи спеціалізовану модель

ukr-detect/ukr-emotions-classifier як основу. Було проведено порівняння трьох підходів:

Еталонний тест: Класифікація на основі вихідних ймовірностей трансформера з оптимальними порогами.

Тільки ознаки розкладу Кунченко: Класифікація з використанням лише 7 згенерованих ознак похибки реконструкції в просторі із порідим елементом.

Гібридна модель: Класифікація на основі об'єднаного вектора ознак (7 ймовірностей + 7 ознак розкладу Кунченка).

Результати, усереднені за метрикою Macro F1-score, представлені в Таблиці 1.

Таблиця 1. Результати класифікації на датасеті EMOBENCH-UA.

Модель	Macro F1-score	Відносний приріст
Еталонний тест (трансформер з порогами)	0.6031	-
Тільки ознаки Кунченко	0.6249	+0.0218
Гібридна модель (трансформер + ознаки Кунченко)	**0.6302**	**+0.0271**

Результати демонструють, що додавання 7 згенерованих ознак до вихідних даних трансформера дозволяє покращити якість класифікації. Бутстреп-тест підтвердив, що цей приріст є статистично значущим, з p-value = 0.01 та 95% довірчим інтервалом для покращення [+0.0051, +0.0479].

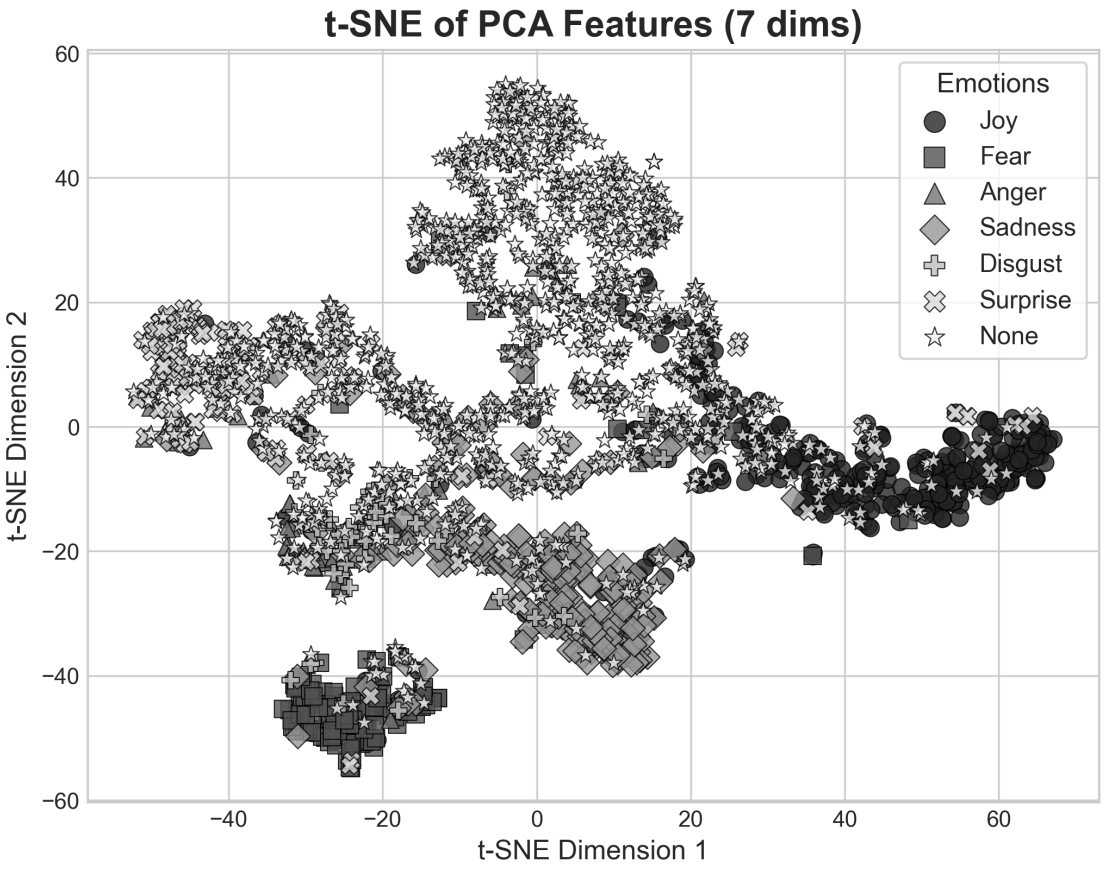
Найбільш неочікуваним і важливим результатом є те, що класифікатор, навчений лише на 7 ознаках Кунченка, не тільки показав високу якість, але й перевершив сильний еталонний тест авторів набору даних на основі самої трансформерної моделі. Це свідчить про високу інформативність «статистико-геометричних» ознак, які описують, наскільки типовим є об'єкт для кожного класу в просторі спеціалізованих векторів.

Для якісної оцінки простору згенерованих ознак був проведений візуальний аналіз за допомогою методу t-SNE, який проєктує багатовимірні дані у двовимірний простір, зберігаючи локальну структуру. На рисунку 1 представлено візуалізацію векторних представлень з тестового набору EMOBENCH-UA для трьох підходів.

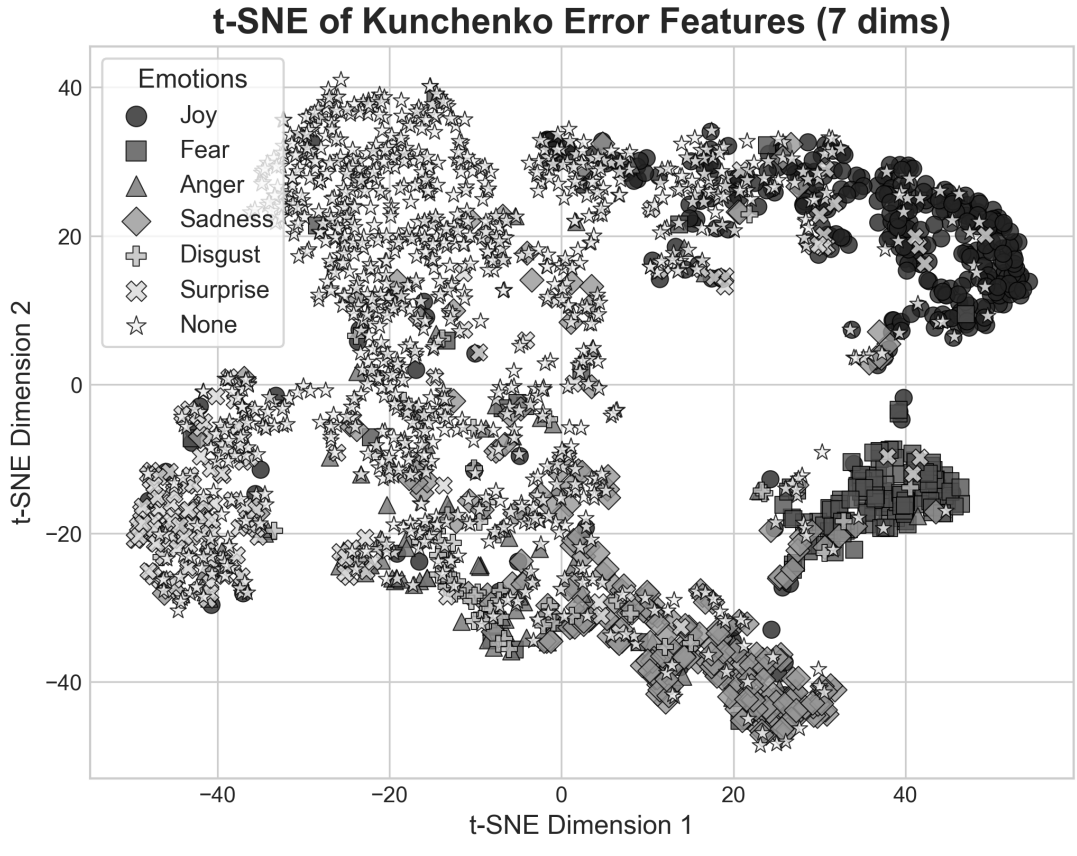
Аналіз візуалізацій дозволяє зробити кілька важливих спостережень. На Рисунку 1(а), що відповідає некерованому методу PCA, видно лише загальну структуру даних. Хоча деякі емоції, як-от «Радість» (Joy) та «Страх» (Fear), формують відносно помітні скупчення, більшість класів сильно перемішані, особливо в центральній частині.

Натомість, на Рисунку 1(б), що демонструє простір ознак, згенерованих методом на основі розкладу Кунченко, кластерна структура є значно чіткішою. Класи емоцій утворюють більш щільні та краще сепаровані групи. Це візуально підтверджує, що наш керований підхід, який враховує приналежність до класу на етапі навчання, створює більш інформативний та семантично розділений простір ознак, що і пояснює вищу якість класифікації.

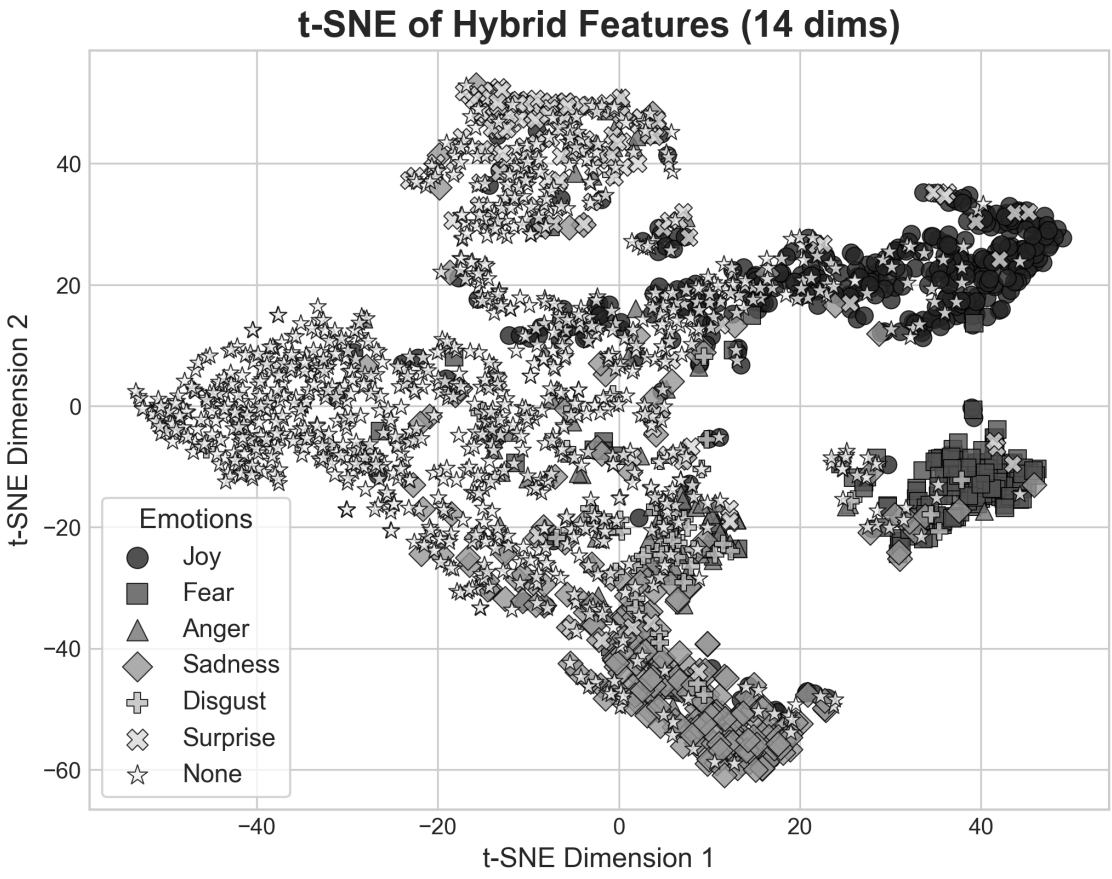
Нарешті, Рисунок 1(в) показує простір гібридних ознак. Він успадковує чітку кластеризацію від ознак на основі розкладу Кунченко, при цьому ще більше «розсуваючи» кластери один від одного. Це свідчить про синергетичний ефект: ознаки PCA додають інформацію про глобальну варіативність, а ознаки на основі розкладу Кунченко — про локальну класо-специфічну структуру, що в сумі дає найкращий результат.


t-SNE of PCA Features (7 dims)
Emotions
Joy
Fear
Anger
Sadness
Disgust
Surprise
None
t-SNE Dimension 2
t-SNE Dimension 1
60
40
20
0
−20
−40
−60
−40
−20
0
20
40
60

а)


t-SNE of Kunchenko Error Features (7 dims)
Emotions
Joy
Fear
Anger
Sadness
Disgust
Surprise
None
t-SNE Dimension 2
t-SNE Dimension 1
40
20
0
−20
−40
−40
−20
0
20
40

б)



в)

Рисунок 1. Візуалізація простору ознак за допомогою t-SNE для датасету EMOBENCH-UA: (а) PCA-ознаки (7 вимірів); (б) Ознаки Кунченка (7 вимірів); (в) Гібридні ознаки (14 вимірів).

3.2. Узагальнювальна здатність та умови застосування методу (EmoEvent)

Другий етап був присвячений перевірці узагальнювальної здатності методу на англомовному датасеті EmoEvent та дослідженню умов його ефективності.

Сценарій А: Без донавчання базової моделі

У цьому сценарії порівнювалися класичні методи та метод на основі розкладу Кунченка, що застосовувалися до «сирих» ембедингів від загальноцільової моделі RoBERTa-base.

Таблиця 2. Результати на EmoEvent без донавчання roberta-base.

Метод	Macro F1-score
TF-IDF + SVM (еталонний тест)	0.3392
PCA (768 -> 50) + SVM	0.3653
Тільки ознаки Кунченко + SVM	**0.1525**

Як видно з таблиці, на неспеціалізованих, «сирих» ембедингах ознаки на основі розкладу Кунченка виявилися неефективними, показавши результат значно нижчий за класичні підходи. Це ключова знахідка, яка визначає межі застосовності методу: для ефективної роботи розкладу Кунченка потрібен попередньо структурований, семантично осмислений простір ембедингів.

Сценарій Б: З донавчанням базової моделі

У цьому сценарії модель RoBERTa-base була донавчена на даних EmoEvent, після

чого до її виходів було застосовано SVM класифікатор. Цей варіант був використаний як еталонний тест. З ним порівнювався гібридний варіант, який додатково включав додаткові ознаки, сформовані на основі рокладу Кунченко. Отримані результати представлені в Таблиці 3.

Таблиця 3. Результати на EmoEvent з донавченою RoBERTa-base.

Метод	Macro F1-score
Еталонний тест (Донавчена RoBERTa) + SVM	0.4602
Гібридна модель (Донавчена RoBERTa + ознаки Кунченко) + SVM	**0.4825**
Приріст	*+0.0223*

Отримані результати свідчать, що після того, як простір ембедингів було спеціалізовано під цільову задачу шляхом донавчання, метод Кунченка продемонстрував свою ефективність, забезпечивши стабільний приріст якості поверх сильного еталонного тесту на основі донавченої моделі RoBERTa. Це підтверджує, що цей метод є потужним «уточнювачем» для вже добре налаштованих моделей, а його ефективність відтворюється на різних мовах.

3.3. Вплив вибору базисних функцій та їх оптимізація

Фінальний етап дослідження був присвячений аналізу впливу вибору базисних функцій на кінцевий результат, оскільки цей вибір є ключовим гіперпараметром методу. Для обох наборів даних було проведено систематичний пошук оптимальних конфігурацій, що включав вибір типу базису, кількості функцій та фінального класифікатора

3.3.1. Оптимізація для українського набору даних (EMOBENCH-UA)

Для українського набору даних було проведено пошук по сітці для параметрів $S = 2..5$ та двох форми базису: дрібно-степеневого або степеневого. Ефективність кожної конфігурації оцінювалася за допомогою двох класифікаторів: Logistic Regression та SVM. Спочатку було встановлено базові результати (бенчмарки), навчивши класифікатори лише на вихідних ймовірностях нейромережі ukr-detect/ukr-emotions-classifier. Вони виявилися практично ідентичними: Macro F1 = 0.6211 для Logistic Regression та 0.6210 для SVM. Результати пошуку по сітці показали, що Logistic Regression дещо ефективніше використовує додаткові ознаки. Найкращий результат Macro F1 = 0.6309 було досягнуто при $S = 4$ і дрібно-степеневому базису. Класифікатор SVM також показав покращення, (Macro F1 = 0.6247), яке було досягнуто при $S = 3$, але степеневому типі базисних функцій.

3.3.2. Оптимізація для англійського набору даних (EmoEvent)

Аналогічний експеримент було проведено для англійського набору даних EmoEvent, де як джерело ембедингів використовувалася донавчена модель RoBERTa-base. Бенчмарком слугував результат самої fine-tuned моделі RoBERTa, який склав Macro F1 = 0.4465. Пошук оптимальної конфігурації показав наступні результати: класифікатор на основі Logistic Regression досягнув найкращого результату Macro F1 = 0.4823 зі степеневим базисом, а SVM показав свою максимальну ефективність Macro F1 = 0.4795 приз використанні дрібно-степеневого базису. Обидва максимуми були отримані при $S = 5$).

3.3.3. Загальні висновки щодо оптимізації

Проведені експерименти доводять, що оптимальний вибір конфігурації гібридної моделі є важливим етапом і залежить від низки факторів:

1. **Геометрія простору ембедингів:** Для української моделі найкраще спрацював

дрібно-степеневий базис, тоді як для англійської RoBERTa — степеневий (у парі з LogReg). Це підтверджує, що архітектура моделі, мова та дані для донавчання формують унікальну геометрію, яка потребує індивідуального підходу.

2. **Вибір класифікатора:** Взаємодія між типом базису та фінальним класифікатором є нетривіальною. Як показав експеримент на EmoEvent, різні класифікатори можуть віддавати перевагу різним типам ознак.

Все це доводить, що оптимальний вибір базисних функцій є важливим гіперпараметром методу. Він залежить від специфічної геометрії конкретного простору ембедингів, яка, у свою чергу, визначається архітектурою моделі, мовою та даними для донавчання. Можливість гнучкого налаштування базису, включно зі створенням комбінованих ознак різноманітних базисних функцій для конкретного набору даних, перетворює запропонований метод на потужний фреймворк для поглибленого аналізу векторних представлень.

4. ВИСНОВКИ

У цьому дослідженні було запропоновано, реалізовано та всебічно перевірено гібридний метод класифікації емоцій, що поєднує потужність сучасних трансформерних моделей з апаратом розкладу в просторі з порідним елементом (просторі Кунченка). На основі проведених експериментів на українському та англійському наборах даних можна зробити наступні висновки:

- **Основна гіпотеза дослідження підтверджена:** Запропонований метод генерації ознак на основі похибки класо-специфічної реконструкції є ефективним інструментом для покращення якості класифікації емоцій. Додавання згенерованих «геометричних» ознак до виходів сильних трансформерних моделей призводить до стабільного та статистично значущого підвищення метрики Macro F1-score.
- **Визначено ключову умову ефективності методу:** Ефективність запропонованого підходу критично залежить від якості та семантичної структурованості вхідного простору ембедингів. Метод демонструє високу ефективність на векторах, отриманих з моделей, що були попередньо донавчені на цільовій задачі. Водночас, він є малоефективним на «сирих», загальноцільових ембедингах. Це позиціонує метод як потужний «уточнювач» (refiner) для вже добре налаштованих систем, а не як універсальний замінник класичних підходів на неструктурованих даних.
- **Доведено узагальнювальну здатність:** Позитивний ефект від застосування методу стабільно відтворюється на різних мовах (українська, англійська), різних базових архітектурах (XLM-RoBERTa, RoBERTa) та різних типах анотації даних (multilabel, multiclass), за умови дотримання вимоги щодо попередньої спеціалізації ембедингів.
- **Встановлено гнучкість методу:** Дослідження показало, що не існує єдиного універсального набору базисних функцій для реконструкції. Оптимальний вибір базису є важливим гіперпараметром, що залежить від специфічної геометрії конкретного простору даних. Це перетворює запропонований підхід з фіксованого алгоритму на гнучкий фреймворк для поглибленого аналізу векторних представлень.

Таким чином, можна зробити загальні висновки, що основним результатом даної робота є успішна адаптація класичних підходів методів статистичного розпізнавання образів до сучасної задачі обробки природної мови та демонстрація практичної цінності гібридного варіанту побудови класифікаційних ознак із використанням розкладу в просторі із порідним елементом (просторі Кунченко).

Результати роботи відкривають кілька перспективних напрямків для подальших досліджень:

- **Застосування до інших NLP-задач:** Перевірка ефективності методу для аналізу

тональності, виявлення токсичності, класифікації намірів та інших задач, де важлива тонка диференціація між класами**.**

- **Автоматизація вибору базису:** Розробка алгоритмів для автоматичного підбору оптимальних базисних функцій. Це потенційно дозволить перейти від поточної фіксованої параметричної форми до більш гнучких підходів для пошуку нетривіальних функціональних залежностей у текстових даних.
- **Інтеграція в нейромережеву архітектуру:** Дослідження можливості вбудовування шару, що обчислює похибку реконструкції, безпосередньо в нейромережу для end-to-end навчання.
- **Оптимізація обчислень:** Пошук шляхів прискорення генерації ознак, наприклад, через попереднє зниження розмірності простору ембедингів за допомогою методів типу PCA.

СПИСОК ЛІТЕРАТУРИ

1 Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports, 2*(7), e12189. https://doi.org/10.1002/eng2.12189

2 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

3 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

4 Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). https://doi.org/10.18653/v1/2020.acl-main.747

5 Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3661). https://doi.org/10.18653/v1/P19-1356

6 Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866. https://doi.org/10.1162/tacl_a_00349

7 Al-Haj, R. A. A., Al-Dossari, H., & Al-Hagery, M. A. (2022). A Hybrid Deep Learning Model for Emotion and Sentiment Analysis Using a Fusion of Transformer and Lexicon Features. *Applied Sciences, 12*(19), 9993. https://doi.org/10.3390/app12199993

8 Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics, 48*(1), 207–218. https://doi.org/10.1162/coli_a_00437

9 Levy, T., Goldman, O., & Tsarfaty, R. (2023). Is probing all you need? Indicator tasks as an alternative to probing embedding spaces. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5243–5254). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.348

10 Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (2nd ed.). Academic Press.

11 Кунченко Ю.П. Полиномы приближения в пространстве с порождающим элементом. – К.: Наук. думка, 2003. – 243 с.

12 Кунченко, Ю. П. Стохастические полиномы. – К.: Наукова думка, 2006. – 275 с
13 Заболотній, С. В. (2009). Статистичне розпізнавання образів на основі розкладу в просторі з порідним елементом. *Вісник Національного університету "Львівська політехніка", № 638: Комп'ютерні науки та інформаційні технології*, 118–123.
14 Zabolotni, S.V. (2014). Application decomposition in space with a generative elements for solving problems of probabilistic diagnostics. *Eastern-European Journal of Enterprise Technologies*, *4*(4(70), 28–35. https://doi.org/10.15587/1729-4061.2014.26195
15 Zabolotnii, S.W., Martynenko, S.S. & Salypa, S.V. Method of Verification of Hypothesis about Mean Value on a Basis of Expansion in a Space with Generating Element. *Radioelectron.Commun.Syst.* 61, 222–229 (2018). https://doi.org/10.3103/S0735272718050060
16 Dementieva, D., Babakov, N., & Fraser, A. (2025). EmoBench-UA: A Benchmark Dataset for Emotion Detection in Ukrainian. *arXiv preprint arXiv:2505.23297*. https://doi.org/10.48550/arXiv.2505.23297
17 Plaza-del-Arco, F. M., Strapparava, C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 1492–1498). European Language Resources Association.

S.V. Zabolotnii, V.I. Hotunov, A.V. Chepynoha

FROM STATISTICAL PATTERN RECOGNITION TO EMOTION ANALYSIS: THE DECOMPOSITION METHOD IN A GENERATING ELEMENT SPACE (KUNCHENKO SPACE) FOR NLP MODELS

Abstract. Emotion recognition in texts is a key task in modern natural language processing, which is dominated by transformer architectures. However, their internal mechanisms remain a 'black box,' and the quality of classification, especially for complex cases, has room for improvement. This paper proposes a new hybrid approach that combines the power of modern NLP models with a deep analysis of their vector representations by adapting the classical method of statistical pattern recognition based on decomposition with a generating element space (Kunchenko space). The method generates a new vector of 'statistical-geometric' features based on the reconstruction error of embeddings by class-specific models.

Experiments on Ukrainian (EMOBENCH-UA) and English (EmoEvent) datasets showed that the presented hybrid approach statistically significantly improves the classification quality (for example, the Macro F1-score on EMOBENCH-UA increased from 0.6031 to 0.6302, p-value=0.01). The study also revealed key conditions for the effectiveness of the method: it is a powerful 'refiner' for models pre-trained on the target task, but ineffective on 'raw', non-specialised embeddings. It has been established that the choice of basis functions for reconstruction is an important hyperparameter that allows the method to be adapted to the specific geometry of the data space.

Keywords: emotion recognition, natural language processing, Kunchenko space, feature generation, hybrid model, embeddings, transformer models.