УДК 004.89:004.912 **DOI:** 10.31673/2412-9070.2026.056711

В. Р. ПЕРЕДЕРА, аспірант;
ORCID: 0009-0001-1397-0079
О. Л. НЕДАШКІВСЬКИЙ, д-р техн. наук, доцент,
ORCID: 0000-0002-1788-4434
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

МЕТОД ПАРАФРАЗ-ІНВАРІАНТНОЇ СТИЛОМЕТРИЧНОЇ ПРОЕКЦІЇ ДЛЯ АТРИБУЦІЇ АВТОРСТВА ТЕКСТІВ

Класичні стилометричні класифікатори авторства, попри високу точність на оригінальних текстах, демонструють значне падіння якості, коли вхідний текст переписано сучасними нейромережевими парафразерами. Опубліковані бенчмарки для атрибуції авторства зазвичай змішують ефект парафразування зі зміною тематики, тому величина чистого падіння точності, спричиненого саме парафразом, залишається маловивченою. У статті досліджено ізольований вплив парафразу на точність стилометричної атрибуції та запропоновано метод PISP-Light (Paraphrase-Invariant Stylometric Projection) – парафраз-інваріантну стилометричну проекцію. Метод застосовує сингулярне розкладення матриці різниць між стилометричними векторами оригінальних і парафразованих тренувальних текстів і будує проектор у підпростір, ортогональний до головних напрямків парафразного шуму. На відміну від парафраз-аугментації, парафрази в PISP слугують діагностичним інструментом для виявлення шумової підструктури, а не додатковими навчальними прикладами. Введено метрику Paraphrase Robustness Drop (PRD) для відтворюваного порівняння методів за стійкістю до парафразу. Експерименти на корпусі The Guardian (13 авторів, 5 випадкових ініціалізацій) показали, що класифікатор на символьних n-грамах зберігає 34 % початкової точності після парафразування, тоді як стилометричний baseline втрачає близько 80 %. У межах стилометричного простору ознак метод PISP із фіксованим порогом пояснюваної дисперсії τ = 0.7 м'яко поліпшує парафразну стійкість (PRD з 0.797 до 0.780) ціною помірного зниження точності на оригінальному тесті. Діагностичний аналіз кривої кумулятивної пояснюваної дисперсії показав, що парафразний шум у стилометричному просторі має лише частково низькорангову структуру, що визначає принципову межу лінійної проекції. Результати мають значення для розробки програмного забезпечення ідентифікації авторства та джерела інформації в мережі Інтернет, стійкого до навмисних перетворень тексту.

Ключові слова: інженерія програмного забезпечення, атрибуція авторства, стилометрія, парафразування, робастність, сингулярне розкладення, нейромережеві перетворення тексту, обчислювальна стилістика, ідентифікація джерела інформації.

Вступ

Атрибуція авторства – задача встановлення особи, яка створила заданий текст, – вже понад десятиліття поєднує класичні стилометричні методи з нейромережевими моделями [1], [2]. Сучасні класифікатори на коротких новинних і блогерських корпусах досягають точності понад 90 % [3], [4]. Однак ці результати майже завжди отримано в «ідеальних» умовах, коли тестові тексти не зазнавали жодних навмисних перетворень. Це обмежує практичну застосовність відповідного програмного забезпечення для виявлення джерела інформації в мережі Інтернет, де текст може бути свідомо змінено для уникнення ідентифікації.

Поширення доступних нейромережевих парафразерів на основі архітектури трансформера [7] (T5, PEGASUS, DIPPER, переписування великими мовними моделями) суттєво змі-

© Передера В. Р., Недашківський О. Л., 2026

нило загрозовий ландшафт. Як показали Sadasivan et al. [8] та Krishna et al. [9], парафразування дозволяє знизити AUROC детекторів машинно-згенерованого тексту з ~0.95 до ~0.55. Macko et al. [10] на бенчмарку 10 обфускаторів × 37 детекторів × 11 мов показали, що практично всі сучасні обфускатори обходять наявні класифікатори. Натомість для класичної атрибуції людських авторів детального вивчення чистого ефекту парафразу майже немає.

У попередніх дослідженнях зі стилометричної атрибуції було показано, що корпуси, які поєднували парафразування та зміну теми, давали падіння точності нижче випадкового рівня, проте не дозволяли відокремити чистий ефект парафразу від тематичного зсуву. У цій статті усунуто цю прогалину: моделі навчаються на оригінальних текстах автора й оцінюються на парафразованих варіантах тих самих текстів (in-domain парафразовий тест).

Запропонований метод PISP-Light ґрунтується на спостереженні, що парафразер систематично змінює певні стилометричні ознаки (часткову заміну функціональних слів, перебудову синтаксичних структур, перерозподіл частиномовних тегів) і слабше – інші. Якщо ці систематичні зміщення займають у просторі стилометричних ознак низькорангову підструктуру, її можна виокремити сингулярним розкладенням матриці різниць між оригіналом і парафразом, а класифікацію виконувати в ортогональному до неї підпросторі. Це принципово відрізняє PISP від методів аугментації даних [11]: парафрази є інструментом діагностики шумової структури, а не додатковими навчальними прикладами.

Постановка проблеми

Програмні засоби ідентифікації авторства та джерела текстової інформації в мережі Інтернет повинні зберігати працездатність в умовах, коли зловмисник навмисно переписує текст автоматичним парафразером. Проте кількісна оцінка того, наскільки саме парафразування (а не супутня зміна тематики чи домену) знижує точність стилометричної атрибуції, у літературі практично відсутня, оскільки наявні бенчмарки змішують ці чинники. Відсутній також відтворюваний показник, який би дозволяв порівнювати методи саме за стійкістю до парафразу.

Звідси випливає наукова задача: ізолювати та виміряти вплив парафразування на точність стилометричної атрибуції авторства, запропонувати відтворювану метрику цього впливу та метод, який підвищує парафразну стійкість стилометричного класифікатора без використання парафразів як додаткових навчальних прикладів. Розв'язання цієї задачі є складовою ширшої проблеми підвищення точності та надійності програмного забезпечення для класифікації та виявлення джерела інформації в мережі Інтернет.

Аналіз останніх досліджень і публікацій

У наших попередніх роботах систематизовано нейромережеві підходи до глибокої стилометрії в задачах визначення авторства – порівняно згорткові, рекурентні та трансформерні архітектури й показано їхню вразливість до стилістичного маскування та навмисного парафразування, що безпосередньо вмотивувало це дослідження [17, 18]. Питання робастності ознак до перетворень тексту досліджувалося переважно в контексті зміни тематики. Stamatatos [3] показав, що символьні n-грами порядку 2–5 зберігають 70–80 % точності класифікації при зміні теми, тоді як слівні ознаки втрачають 15–25 %; у наших експериментах цей висновок підтверджено й поширено на випадок парафразування, а не лише зміни теми. Окремий напрям присвячено парафразним атакам на детектори згенерованого тексту: Sadasivan et al. [8] та Krishna et al. [9] продемонстрували, що парафразери типу DIPPER знижують ефективність таких детекторів до близького до випадкового рівня, а Macko et al. [10] на бенчмарку з 10 обфускаторів, 37 детекторів та 11 мов підтвердили масштабність проблеми.

Серед методів протидії обфускації Zhai et al. [12] застосовують змагальне навчання з CNN/LSTM проти правилоорієнтованих обфускаторів та Mutant-X, знижуючи ефективність обфускації з 20–30 % до 5–10 %, проте визнають, що сучасні нейропарафразери в їхній експериментальній постановці не оцінювалися. Boenninghoff et al. [6], [13], переможці змагання PAN-2019, поєднують семантичні ембединги BERT [5] із синтаксичними частиномовними n-грамами через сіамську мережу для верифікації авторства (AUC ~0.93), що, однак, є іншою за-

дачею та архітектурою порівняно із закритою атрибуцією в умовах парафразу. Концептуально близькі роботи з виявлення підпросторів завад на ембедингах мовних моделей застосовують контрастне навчання для атрибуції джерельної моделі, тобто оперують іншим простором ознак (глибокі ембединги, а не інтерпретовний стилометричний вектор) та іншим механізмом побудови підпростору (контрастне навчання, а не сингулярне розкладення в замкнутій формі).

Отже, раніше невирішеною частиною загальної задачі залишається побудова лінійної проекції парафразного шуму в інтерпретовному стилометричному просторі для закритої атрибуції авторства, а також відтворюване вимірювання чистого ефекту парафразу. Саме цьому присвячено цю статтю.

Мета і задачі дослідження

Мета статті – кількісно оцінити ізольований вплив парафразування на точність стилометричної атрибуції авторства та запропонувати метод і метрику, що підвищують відтворюваність і робастність відповідного програмного забезпечення.

Для досягнення мети поставлено такі задачі:

– розробити метод PISP-Light із замкнутою формулою побудови проектора через сингулярне розкладення матриці різниць стилометричних векторів пар «оригінал – парафраз»;

– запропонувати метрику Paraphrase Robustness Drop (PRD) для систематичного порівняння методів за стійкістю до парафразу;

– експериментально порівняти три класифікатори (символьний TF-IDF, стилометричний та PISP-Light) на корпусі The Guardian з п'ятьма випадковими ініціалізаціями та статистичними тестами значущості;

– виконати діагностичний аналіз кривої кумулятивної пояснюваної дисперсії, що пояснює межі лінійної проекції в стилометричному просторі.

Результати дослідження

Стилометричний вектор ознак

Для кожного тексту x обчислюється вектор $\varphi(x) \in \mathbb{R}\text{^}d$ ($d = 120$) з п'яти груп ознак: (1) метрики довжини (3) – середня довжина речення у символах і словах, середня довжина слова в символах; (2) лексичне різноманіття (3) – type-token ratio, частка hapax legomena, кількість речень; (3) розподіл пунктуації (6) – частоти знаків « , . ? ! ; : » на 1000 символів; (4) частиномовний розподіл (8) – відносні частки тегів NOUN, VERB, ADJ, ADV, PRON, PROPN, AUX, DET (через бібліотеку spaCy [16], модель en_core_web_sm); (5) функціональні слова (100) – відносні частоти 100 найуживаніших службових слів англійської. Вектори стандартизуються (StandardScaler, навчений на оригінальних тренувальних даних).

Метод PISP-Light

Нехай $\{(x_i, y_i)\}$, $i = 1..n$ – тренувальний набір. Для кожного x_i генерується парафраз $\tilde{x}_i = T5(x_i)$ моделлю humarin/chatgpt_paraphraser_on_T5_base (beam = 4, температура 1.5, top-p = 0.95). Парафрази не додаються до тренувального набору; вони використовуються для побудови матриці різниць N, кожен рядок якої – вектор зсуву стилометричних ознак, спричиненого парафразуванням конкретного тренувального тексту:

$$N = [\, \varphi(\tilde{x}_1) - \varphi(x_1)\,;\ \ldots\ ;\ \varphi(\tilde{x}_n) - \varphi(x_n)\,] \in \mathbb{R}\text{^}(n \times d) \quad (1)$$

Стовпці N центруються, після чого виконується тонке сингулярне розкладення:

$$N = U \cdot \Sigma \cdot V^{T},\ \Sigma = diag(\sigma_1, \ldots, \sigma_d),\ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0 \quad (2)$$

де $V = [v_1, \ldots, v_d]$ – матриця правих сингулярних векторів. Перші k векторів $V_k = [v_1, \ldots, v_k]$ задають підпростір, у якому зосереджено максимальну частку дисперсії парафразного шуму. Розмірність k обирають за порогом кумулятивної пояснюваної дисперсії τ:

$$k = min\{\, k :\ (\Sigma_{j=1}^{k} \sigma_j^2) / (\Sigma_{j=1}^{d} \sigma_j^2) \geq \tau \,\},\ \tau \in [0,1] \quad (3)$$

Проектор у «робастний» підпростір (ортогональний до напрямків шуму):

$$P_R = I_d - V_k \cdot V_k^{T} \in \mathbb{R}\text{^}(d \times d) \quad (4)$$

Робастна репрезентація тексту:

$$\varphi_R(x) = P_R \cdot S(\varphi(x)) \quad (5)$$

де S – стандартизуючий перетворювач. Класифікатор (логістична регресія, розв'язувач saga, C = 1.0; реалізація scikit-learn [15]) навчається на $\{(\varphi_R(x_i), y_i)\}$, тобто лише на оригінальних тренувальних прикладах у спроєктованому просторі. Метод є замкнутим: окрім фінальної логістичної регресії, він не потребує навчання нейромереж.

Метрика парафразної стійкості PRD

Нехай Acc_T1 – точність моделі на оригінальному тестовому розбитті, Acc_T2 – на парафразованій версії того ж розбиття (по одному парафразу на тестовий текст, мітки збережено). Тоді метрику Paraphrase Robustness Drop визначено як:

$$PRD = (Acc_T1 - Acc_T2) / Acc_T1 \quad (6)$$

Менше значення – краще; PRD = 0 означає повну стійкість моделі до парафразу. Довірчі інтервали оцінюються за формулою $\bar{x} \pm 1.96 \cdot s / \sqrt{n}, n = 5$.

Дані та протокол експерименту

Експерименти проведено на корпусі The Guardian, що містить 439 текстів 13 авторів-колумністів (у середньому ~1040 слів на текст, медіана 1134; від 22 до 44 текстів на автора) зі стратифікованим розбиттям train/test у співвідношенні 80/20. Для парафразування використано реалізовану в бібліотеці Transformers [14] модель humarin/ chatgpt_paraphraser_on_T5_base (T5-base, до-навчену на парафразах ChatGPT) з такими параметрами генерації: променевий пошук із 4 променями, температура 1.5, top-p = 0.95, top-k =50, увімкнене семплування. Кожен текст корпусу парафразовано одноразово, а парафразовані розбиття побудовано тим самим train_test_split із тими ж початковими станами, що гарантує порядкову відповідність між оригінальним і парафразованим train/test.

Порівнювалися три класифікатори. Перший (B1) – TF-IDF на символьних n-грамах (char_wb, порядок 2–5) з логістичною регресією. Другий (B2) – стилометричний вектор $\varphi(x)$ зі стандартизацією та логістичною регресією без проекції. Третій (B4) – запропонований PISP-Light, тобто B2 з проектором P_R. Для B4 розглянуто два режими вибору порога: B4-val, де τ обирається з множини $\{0.5, 0.7, 0.8, 0.9\}$ за точністю на 10 % валідаційному виокремленні з train, та B4-fixed із фіксованим τ для аналізу чутливості.

Сітка експериментів охоплює 3 моделі × 2 тестові умови (T1 – оригінал, T2 – парафраз) × 5 початкових станів (42, 123, 456, 789, 1024), що дає 30 базових експериментів, а також аналіз чутливості для B4-fixed (4 значення τ × 2 умови × 5 станів, тобто 40 експериментів). Оцінювалися accuracy, F1-macro та PRD з 95 % довірчими інтервалами; для пари B2 ↔ B4-val на парафразованому тесті застосовано тест МакНемара. Програмне забезпечення реалізовано за модульною конвеєрною архітектурою; порядкову відповідність розбиттів забезпечує детермінізм train_test_split за фіксованим random_state.

Результати експериментів

Агреговані по 5 seeds результати наведено в табл. 1 та на рис. 1 і рис. 2.

Таблиця 1

Точність на оригінальному (T1) та парафразованому (T2) тесті, PRD з 95 % довірчими інтервалами (корпус Guardian, 13 авторів, 5 seeds)

Модель	Acc_T1 ± CI	Acc_T2 ± CI	PRD ± CI
B1 (TF-IDF char-LR)	0.755 ± 0.023	0.257 ± 0.024	0.660 ± 0.028
B2 (стилометрія)	0.793 ± 0.019	0.161 ± 0.023	0.797 ± 0.028
B4-val (PISP, τ=val)	0.789 ± 0.016	0.159 ± 0.032	0.799 ± 0.036
B4-fixed (τ=0.7)	0.777 ± 0.034	0.171 ± 0.028	0.780 ± 0.037

Найнижчий PRD має B1 на символьних n-грамах (0.660) – це підтверджує робастність символьних ознак, відому з [3] для зміни теми, і поширює її на випадок парафразу (рис. 1). У межах стилометричного простору φ(x) метод PISP з τ = 0.7 м'яко поліпшує PRD з 0.797 до 0.780, проте довірчі інтервали майже повністю перекриваються. Тест МакНемара між B2 і B4-val на T2 дає p-значення в діапазоні [0.38; 1.00] (у середньому 0.80 за 5 seeds), тобто різниця статистично не значуща. Це означає, що автоматичний вибір τ за оригінальною валідацією не

дає надійного виграшу, і робастність символьних ознак залишається недосяжною для лінійної проекції в стилометричному просторі.

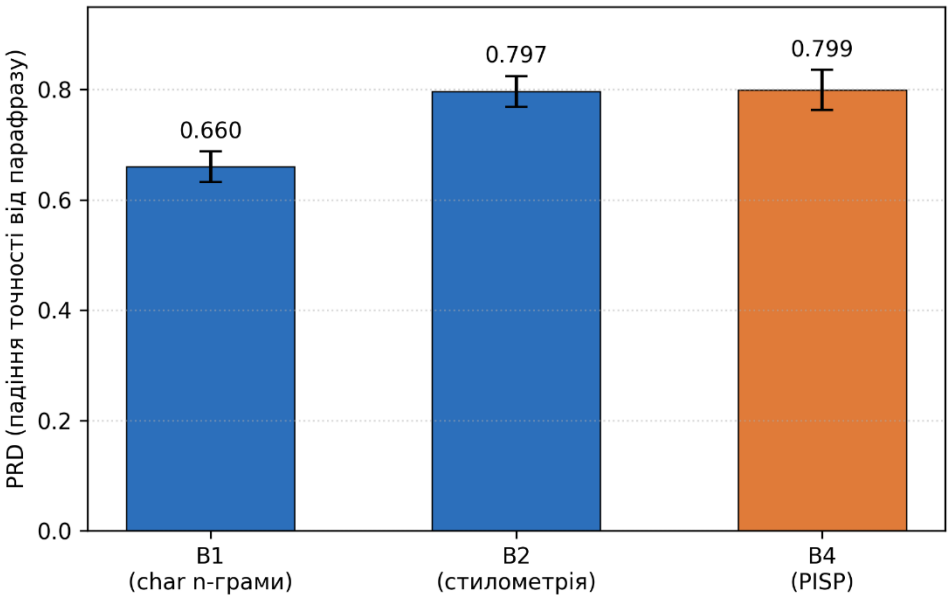


Рис. 1. PRD для кожної моделі з 95 % довірчими інтервалами (корпус Guardian, 5 seeds)

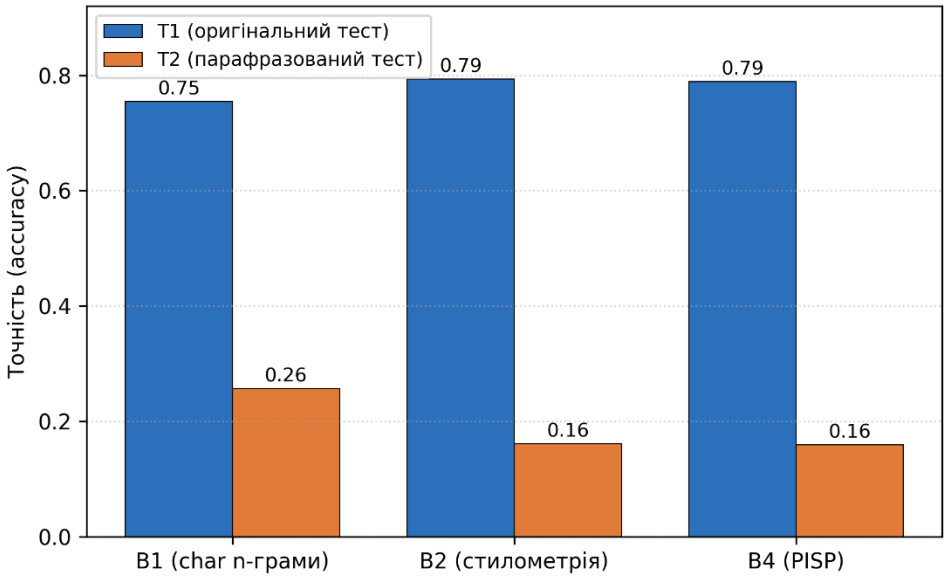


Рис. 2. Точність на оригінальному (T1) та парафразованому (T2) тесті для кожної моделі

Аналіз чутливості до порога τ

Поведінку PISP із фіксованим τ наведено в табл. 2 (k – середня кількість відкинутих сингулярних компонент).

Таблиця 2

PISP-Light з фіксованим порогом τ (корпус Guardian, 5 seeds)

τ	k (сер.)	Acc_T1 ± CI	Acc_T2 ± CI	PRD ± CI
0.5	6.0	0.786 ± 0.017	0.157 ± 0.028	0.801 ± 0.033
0.7	19.0	0.777 ± 0.034	0.171 ± 0.028	0.780 ± 0.037
0.9	48.8	0.711 ± 0.041	0.159 ± 0.028	0.774 ± 0.046
0.95	64.8	0.616 ± 0.049	0.159 ± 0.035	0.738 ± 0.066

Спостерігається чіткий компроміс: збільшення k (агресивніша проекція) зменшує PRD, але водночас знижує Acc_T1. Точка τ = 0.7 (відкидається у середньому 19 зі 120 компонент) є компромісною: PRD знижено на 0.021 від baseline B2, а Acc_T1 – лише на 0.016. Жорсткіша

проекція ($\tau = 0.95$) дає менший PRD (0.738) ціною значного падіння Acc_T1 (0.616), що робить її непрактичною.

Валідаційний критерій, що максимізує точність на 10 % оригінального train (~35 текстів), систематично обирає $\tau = 0.5$ ($k = 3..13$) у всіх 5 seeds. Причина: валідаційна вибірка не містить парафразів, тож метрика винагороджує точність на оригіналі й штрафує будь-яке «сильне» відкидання компонент, навіть коли воно поліпшує T2. Це самостійна методологічна знахідка: для парафраз-стійкої атрибуції стандартний критерій вибору гіперпараметрів непридатний; потрібен критерій, що враховує парафразну стійкість (наприклад, парафраз-доповнена валідація).

Крива пояснюваної дисперсії

Перші три сингулярні компоненти матриці N покривають у середньому ~45 % її дисперсії (від 29 % до 53 % залежно від seed), після чого частки спадають повільно: для досягнення $\tau = 0.9$ потрібно в середньому ~49 компонент, для $\tau = 0.95$ – ~65 (рис. 2). Це означає, що парафразний шум у стилометричному просторі має лише частково низькорангову структуру: значна частина варіації рівномірно «розмазана» по багатьох напрямках. Звідси випливає принципова межа лінійної проекції: суттєву частку парафразного шуму неможливо усунути малою кількістю компонент без втрати інформації про авторство.

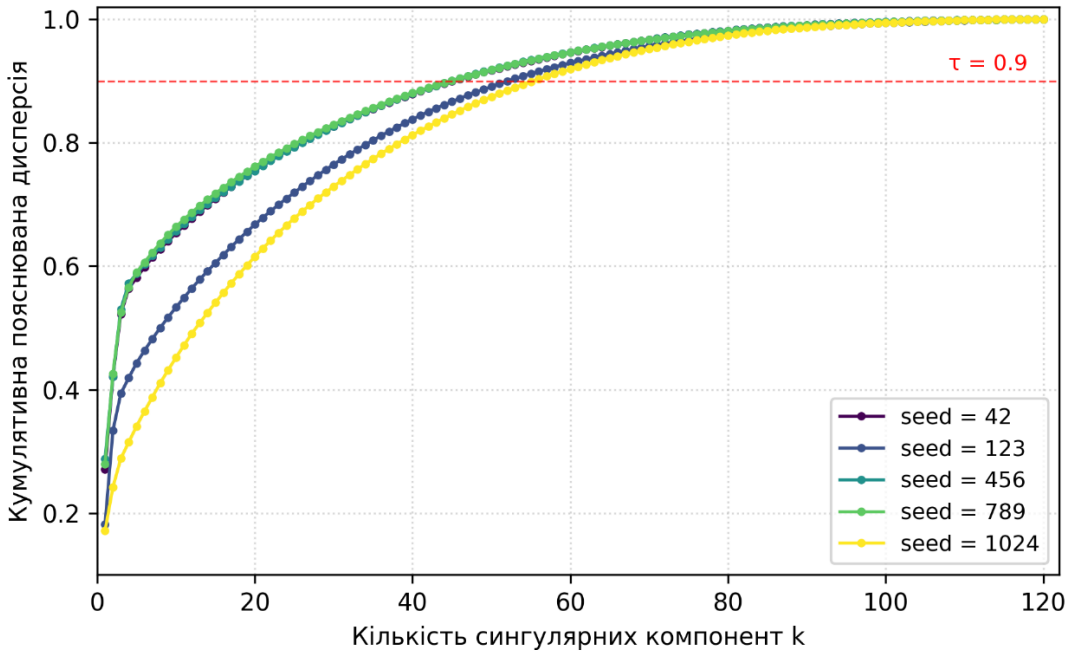


Рис. 3. Кумулятивна пояснювана дисперсія матриці парафразного шуму N (5 seeds)

Обговорення

Усі три моделі демонструють великі PRD (0.66–0.80), що ставить під сумнів практичну застосовність стилометричних класифікаторів в умовах, коли тексти можуть бути парафразовані, і вмотивовує розробку парафраз-стійких програмних засобів ідентифікації джерела інформації. При цьому B1 на символьних n-грамах виявився єдиною моделлю, що зберігає понад 25 % оригінальної точності після парафразу: ймовірно, парафразер змінює лексичні шари сильніше, ніж низькорівневі літерні поєднання, які відображають морфологічні та орфографічні преференції автора.

Стилометричний вектор $\varphi(x)$, навпаки, виявляється чутливішим до парафразу, ніж символьні n-грами (PRD 0.797 проти 0.660), усупереч інтуїції щодо робастності ручних ознак. Ймовірне пояснення полягає в тому, що сучасні нейропарафразери навчилися «усереднювати» саме ті ознаки, які входять до канонічного стилометричного вектора, – функціональні слова, частиномовний розподіл, довжину речень.

Отримані результати слід інтерпретувати з урахуванням низки обмежень. Дослідження виконано на одному датасеті (Guardian, 13 авторів), оскільки генерація парафраз для більших корпусів обмежена обчислювальними ресурсами, і лише для англомовних текстів. Використ-

тано один парафразер (T5-base), тож крива стійкості проти інших систем (DIPPER, PEGASUS, переписування великими мовними моделями) залишається відкритим питанням. Запропонована проекція є лінійною, тоді як нелінійний еквівалент міг би вловити «розмазану» складову шуму. Нарешті, розглянуто закриту атрибуцію з 13 авторами, а перенесення на верифікацію авторства та відкриту атрибуцію потребує окремого вивчення.

Висновки та перспективи подальших досліджень

1. Запропоновано метод PISP-Light – парафраз-інваріантну стилометричну проекцію, що в замкнутій формі виокремлює напрямки парафразного шуму через сингулярне розкладення матриці різниць стилометричних векторів пар «оригінал – парафраз» і класифікує в ортогональному до них підпросторі. Парафрази слугують інструментом діагностики, а не аугментованими прикладами, що принципово відрізняє підхід від аугментації даних.

2. Введено метрику Paraphrase Robustness Drop (PRD) для відтворюваного порівняння методів за стійкістю до парафразу in-domain. Експерименти на корпусі Guardian показали: символьні n-грами (B1) – помітно найстійкіша модель (PRD = 0.660); у межах стилометричного простору $\varphi(x)$ метод PISP із фіксованим $\tau = 0.7$ м'яко поліпшує робастність над baseline (PRD з 0.797 до 0.780) ціною помірного зниження точності на оригіналі. Стандартний критерій вибору τ за точністю на оригінальній валідації систематично недооцінює оптимальне значення – самостійна методологічна знахідка.

3. Діагностичний аналіз показав, що парафразний шум у стилометричному просторі має лише частково низькорангову структуру (перші три компоненти покривають у середньому ~45 % дисперсії), що визначає принципову межу лінійної проекції.

Перспективи подальших досліджень: (1) Soft-PISP із ваговою (а не бінарною) проекцією; (2) множинні парафразери (DIPPER, PEGASUS, переписування великими мовними моделями) для побудови повної кривої стійкості; (3) парафраз-врахований критерій вибору гіперпараметрів на парах «оригінал – парафраз»; (4) нелінійні розширення (kernel PISP, автокодувальне звуження, контрастний підпростір) для роботи з «розмазаною» складовою шуму; (5) крос-лінгвальні експерименти на українському корпусі; (6) перенесення на верифікацію авторства через сіамські архітектури.

Внесок авторів

В. Р. ПЕРЕДЕРА – концептуалізація, розробка методу та програмного забезпечення, проведення експериментів, аналіз даних, написання тексту; О. Л. НЕДАШКІВСЬКИЙ – наукове керівництво, постановка задачі, аналіз і верифікація результатів, редагування рукопису.

Декларація про штучний інтелект

Інструменти на основі штучного інтелекту використовувалися для допоміжного мовностилістичного редагування та оформлення рукопису. Нейромережева модель парафразування T5 є компонентом досліджуваного методу та об'єктом дослідження. Усі наукові результати, програмна реалізація, експерименти, їх аналіз та висновки отримані й перевірені авторами, які несуть повну відповідальність за зміст статті.

Конфлікт інтересів

Автори заявляють про відсутність конфлікту інтересів та підтверджують, що під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б бути розцінені як такі, що здатні вплинути на результати дослідження або їх інтерпретацію. Роботу виконано відповідно до принципів академічної доброчесності та етичних норм проведення наукових досліджень.

Список використаної літератури

1. Stamatatos E. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology. 2009. Vol. 60, № 3. P. 538–556. https://doi.org/10.1002/asi.21001

2. Koppel M., Schler J., Argamon S. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology. 2009. Vol. 60, № 1. P. 9–26. https://doi.org/10.1002/asi.20961

3. Stamatatos E. On the robustness of authorship attribution based on character n-gram features. Journal of Law & Policy. 2013. Vol. 21, № 2. P. 421–439.

4. Fabien M., Villatoro-Tello E., Motlicek P., Shankar B. BertAA: BERT fine-tuning for authorship attribution. Proceedings of the 17th International Conference on Natural Language Processing (ICON). 2020. P. 127–137.

5. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. 2019. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423

6. Boenninghoff B., Hessler S., Kolossa D., Nickel R. M. Explainable authorship verification in social media via attention-based similarity learning. Proceedings of IEEE BigData. 2019. P. 36–45. https://doi.org/10.1109/BigData47090.2019.9005650

7. Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017. Vol. 30. https://doi.org/10.48550/arXiv.1706.03762

8. Sadasivan V. S., Kumar A., Balasubramanian S., Wang W., Feizi S. Can AI-generated text be reliably detected? arXiv preprint. 2023. https://doi.org/10.48550/arXiv.2303.11156

9. Krishna K., Song Y., Karpinska M., Wieting J., Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. Advances in Neural Information Processing Systems. 2023. Vol. 36. https://doi.org/10.48550/arXiv.2303.13408

10. Macko D., Moro R., Uchendu A. et al. Authorship obfuscation in multilingual machine-generated text detection. arXiv preprint. 2024. https://doi.org/10.48550/arXiv.2401.07867

11. Wei J., Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of EMNLP-IJCNLP. 2019. P. 6382–6388. https://doi.org/10.18653/v1/D19-1670

12. Zhai R., Singh J., Wei Z., Vaidya J., Atallah M. Adversarial authorship attribution for deobfuscation. Proceedings of ACL. 2022. P. 7372–7384. https://doi.org/10.18653/v1/2022.acl-long.509

13. Boenninghoff B., Nickel R. M., Zeiler S., Kolossa D. Similarity learning for authorship verification in social media. Proceedings of IEEE ICASSP. 2019. P. 2457–2461. https://doi.org/10.1109/ICASSP.2019.8683405

14. Wolf T., Debut L., Sanh V. et al. Transformers: State-of-the-art natural language processing. Proceedings of EMNLP: System Demonstrations. 2020. P. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

15. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

16. Honnibal M., Montani I., Van Landeghem S., Boyd A. spaCy: Industrial-strength natural language processing in Python. Zenodo. 2020. https://doi.org/10.5281/zenodo.1212303

V. Peredera, O. Nedashkivskyi

PARAPHRASE-INVARIANT STYLOMETRIC PROJECTION FOR AUTHORSHIP ATTRIBUTION OF TEXTS

Classical stylometric authorship classifiers, while highly accurate on original texts, suffer substantial accuracy drops when the input is rewritten by modern neural paraphrasers. Published authorship-attribution benchmarks typically conflate paraphrasing with topic shift, leaving the isolated effect of paraphrasing on attribution accuracy poorly quantified. This article isolates and measures that effect and proposes PISP-Light (Paraphrase-Invariant Stylometric Projection). The method applies singular value decomposition to the matrix of differences between the stylometric feature vectors of original and paraphrased training texts and projects the features into the subspace orthogonal to the dominant paraphrase-noise directions. In contrast to paraphrase augmentation, PISP uses paraphrases as a diagnostic instrument for discovering the nuisance subspace rather than as additional

training samples. A reproducible Paraphrase Robustness Drop (PRD) metric is introduced to compare methods under paraphrase attacks. Experiments on the Guardian corpus (13 authors, 5 random seeds) show that a character n-gram classifier preserves 34 % of its original accuracy after paraphrasing, whereas a stylometric baseline loses about 80 %. Within the stylometric feature space, PISP with a fixed explained-variance threshold $\tau = 0.7$ mildly improves robustness (PRD from 0.797 to 0.780) at the cost of a moderate decrease in original-test accuracy. A diagnostic analysis of the cumulative explained-variance curve shows that paraphrase noise in the stylometric space has only a partly low-rank structure, which sets a principled limit on linear projection. The results are relevant to developing software for authorship and information-source identification on the Internet that is robust to deliberate text transformations.

Keywords: software engineering**,** authorship attribution, stylometry, paraphrasing, robustness, singular value decomposition, neural text rewriting, computational stylistics, information-source identification.

Надійшла до редакції: 16.09.2026
Прийнята до друку: 06.10.2026
Опубліковано: 30.10.2026
© 2026 Передера В. Р., Недашківський О. Л.
Цей матеріал ліцензовано за умовами CC BY 4.0.https://creativecommons.org/licenses/by/4.0/