DOI: 10.20998/2079-0023.2026.01.16
УДК 004.89:519.76

Т. В. ПЕТРИШАК, аспірант кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», м. Львів, Україна; e-mail: taras.v.petryshak@lpnu.ua; ORCID: https://orcid.org/0009-0006-6296-3867
В. А. ВИСОЦЬКА, доктор технічних наук, професор кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», м. Львів, Україна; e-mail: Victoria.A.Vysotska@lpnu.ua; ORCID: https://orcid.org/0000-0001-6417-3689

ОЦІНЮВАННЯ ЗДАТНОСТІ МОДЕЛЕЙ ВИЯВЛЕННЯ AI-ЗГЕНЕРОВАНИХ ТЕКСТІВ ДО УЗАГАЛЬНЕННЯ В УМОВАХ НЕВІДОМОГО ГЕНЕРАТОРА

Багато детекторів AI-згенерованих текстів демонструють високі показники якості на вибірках, сформованих у межах типових протоколів оцінювання. Зокрема, класичні моделі, що використовують стилометричні ознаки, такі як довжина тексту, пунктуаційні патерни та узагальнені показники формальності, здатні ефективно виявляти статистичні закономірності генерації. Проте їхня ефективність суттєво знижується у випадках, коли з'являються тексти, створені невідомими генераторами. У таких умовах розподіл ознак змінюється, що призводить до погіршення якості класифікації, насамперед через зростання кількості хибно негативних помилок. У роботі досліджено здатність моделей до узагальнення в умовах появи невідомого генератора. Порівняння проводиться між класичними стилометричними моделями та трансформерними підходами з використанням експериментального протоколу LOGO (Leave-One-Generator-Out). Розглядається задача бінарної класифікації текстів у двох доменах (Reddit, Wikipedia) із використанням трьох генераторів: ChatGPT, Davinci та Dolly. До класичних моделей віднесено Random Forest і Gradient Boosting, тоді як трансформерні підходи представлені моделями DistilBERT і RoBERTa. Якість моделей оцінювалася за метриками Accuracy, Precision, Recall, F1 та Macro-F1 з подальшим усередненням результатів за кількома ініціалізаціями. Отримані результати свідчать, що трансформерні моделі демонструють вищу здатність до узагальнення на даних, згенерованих невідомими моделями. Водночас стилометричні підходи демонструють істотне погіршення якості, зокрема залежно від домену та довжини тексту. Аналіз помилок показує, що ключовим фактором зниження ефективності є зростання кількості хибно негативних класифікацій. Додатковий аналіз важливості ознак підтверджує, що класичні моделі значною мірою залежать від поверхневих характеристик тексту, які не забезпечують стабільної узагальнювальної здатності. Отже, результати дослідження підкреслюють необхідність оцінювання детекторів у межах протоколу LOGO для досягнення надійної стійкості до появи нових генераторів.

Ключові слова: виявлення згенерованих текстів, здатність до узагальнення, невідомий генератор, стилометричні ознаки, трансформерні моделі, класифікація текстів.

Вступ. Більшість сучасних детекторів AI-згенерованих текстів демонструють високі показники точності під час оцінювання на бенчмарках [1]. Класичні моделі машинного навчання, що базуються на стилометричних ознаках, зокрема довжині речень, частотності n-грам, пунктуаційних патернах та узагальнених показниках формальності, здатні ефективно виявляти статистичні закономірності машинної генерації за умови роботи в межах відомого розподілу даних [2]. Це пояснюється тим, що великі мовні моделі під час генерації формують відносно стабільні структурні патерни.

Водночас на практиці ефективність таких підходів суттєво знижується у разі появи нового, невідомого алгоритму генерації. У таких умовах поверхневі сигнали, на які спираються стилометричні методи, змінюються, оскільки різні моделі формують тексти з відмінними характеристиками довжини, пунктуації та словникового складу. Це призводить до погіршення якості детекції, зокрема через зростання кількості хибно негативних помилок, коли згенерований текст класифікується як людський. Зазначена проблема пов'язана з явищем генераторного зсуву, яке у ширшому контексті відповідає задачі узагальнення моделей за межами навчального розподілу даних. Для її коректного дослідження необхідні підходи, що забезпечують ізоляцію джерела генерації під час оцінювання.

У цій роботі розглядається задача оцінювання здатності моделей до узагальнення в умовах появи невідомого генератора з використанням протоколу LOGO (Leave-One-Generator-Out) [3].

Методи дослідження. Методологія дослідження базується на формуванні контрольованих умов тестування, за яких генератор тексту, що використовується на етапі оцінювання, повністю відсутній у навчальній вибірці та є невідомим для моделі детектора. Такий підхід дає змогу коректно оцінити здатність моделей до узагальнення в умовах зміни джерела генерації. У межах дослідження розв'язується задача бінарної класифікації текстів, у якій тексти людського авторства позначаються міткою Human (0), а тексти, згенеровані мовними моделями, міткою Machine (1).

Експеримент виконано на підмножині розмічених даних мультидоменного бенчмарку M4 [1]. Для аналізу обрано два домени, що суттєво відрізняються за стилем: Reddit (неформальні дискусії) та Wikipedia (формальний, структурований текст). Кількісні характеристики тестових наборів даних наведено в табл. 1.

Таблиця 1 – Кількісні характеристики тестових вибірок

Сценарій	Обсяг	Reddit	Wiki	Людські	Машинні
logo_chatgpt	3600	1842	1758	1800	1800
logo_davinci	3600	1842	1758	1800	1800
logo_dolly	3514	1842	1672	1800	1714

Як видно з табл. 1, обсяг тестової вибірки для кожного зі сценаріїв оцінювання складав близько 3500–3600 унікальних текстів. Розподіл даних за доменами був відносно рівномірним, із незначною перевагою текстів із платформи Reddit. Для забезпе-

чення чистоти експерименту кількість текстів негативного класу (Human) була суворо зафіксована на рівні 1800 екземплярів для кожного сценарію тестування, а відхилення в обсязі класу Machine для моделі Dolly відображає природний розподіл даних після етапу очищення.

Загальна архітектура експерименту та основні етапи дослідження представлені на рис. 1. Процес реалізації включає чотири взаємопов'язані логічні блоки, а саме підготовку вхідних даних, формування та

Для класифікації було порівняно дві групи алгоритмів. Перша група охоплює трансформерні моделі DistilBERT і RoBERTa [7, 8], що використовують контекстне донавчання з обмеженням довжини вхідної послідовності до 512 токенів. До другої групи віднесено класичні стилометричні ансамблі Random Forest (RF) та Gradient Boosting (GB) [9, 10]. На відміну від трансформерних моделей, класичні підходи навчалися на спеціально сформованому векторі ознак, отриманому на етапі конструювання ознак (feature

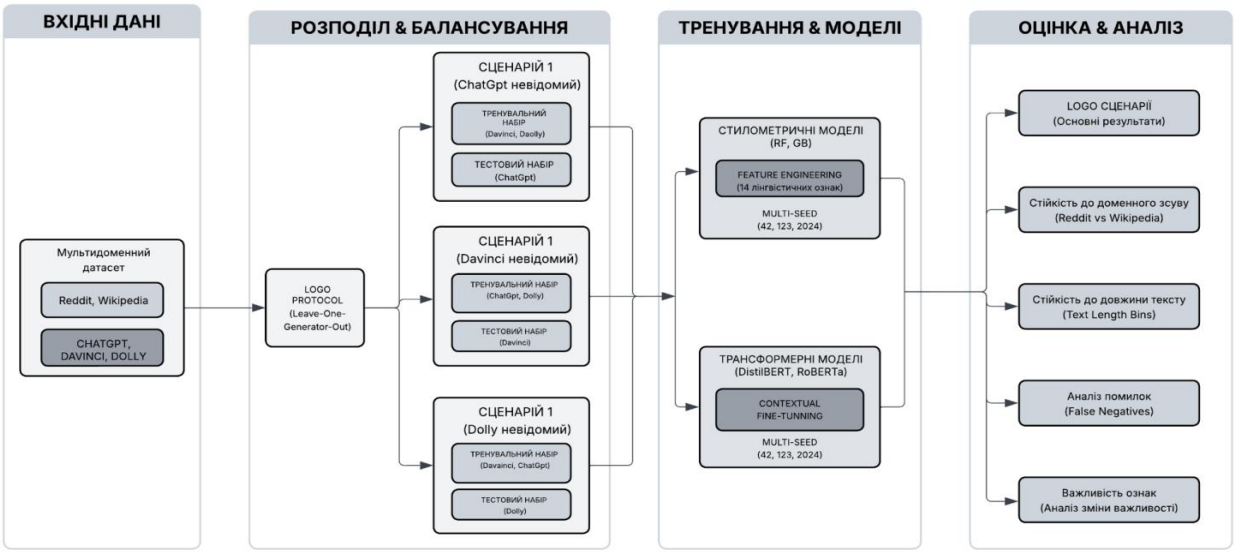

Рис. 1. Архітектура експерименту та етапи дослідження

балансування вибірок, навчання моделей, а також комплексне оцінювання їхньої стійкості та узагальнювальної здатності.

У дослідженні використано тексти, згенеровані трьома мовними моделями з відмінними архітектурними характеристиками: ChatGPT, Davinci та Dolly [4]. Для усунення дублювань застосовувалося хешування, а для запобігання витоку даних розподіл вибірок виконувався методом GroupShuffleSplit із групуванням за ідентифікатором group_id, що поєднує домен і джерело [5].

Ключовим елементом оцінювання є протокол LOGO. У межах експерименту реалізовано три незалежні сценарії (logo_dolly, logo_chatgpt, logo_davinci), у кожному з яких один генератор повністю виключається з навчальної та валідаційної вибірок і з'являється лише на етапі тестування. Для забезпечення коректності навчання тренувальні вибірки (train) були збалансовані за класами за допомогою методів undersampling [6], тоді як тестові вибірки (test) зберігали природний розподіл. Щоб гарантувати стабільність результатів і відокремити вплив випадкового розбиття від впливу ініціалізації моделі, розподіл даних (split) фіксувався єдиним параметром split_seed (42), а навчання моделей проводилося багаторазово з різними параметрами ініціалізації train_seed (42, 123, 2024). Усі кінцеві метрики агрегувалися у форматі середнього значення та стандартного відхилення.

engineering), який містить 14 стилометричних характеристик [11]. Ці ознаки описують базові поверхневі лінгвістичні властивості текстів і наведені в табл. 2.

Таблиця 2 – Стилометричні ознаки

Назва ознаки	Коротке пояснення
simpsons_d	Лексичне різноманіття (індекс Сімпсона)
yules_k	Лексичне багатство (індекс Юла)
type_token_ratio	Відношення типів до токенів
word_count	Загальна кількість слів у тексті
sentence_count	Загальна кількість речень у тексті
avg_word_length	Середня довжина слова у символах
avg_sentence_length	Середня кількість слів у реченні
punctuation_ratio	Частка пунктуації
function_word_ratio	Частка службових слів
noun_ratio	Відносна частка іменників у тексті
verb_ratio	Відносна частка дієслів у тексті
pronoun_ratio	Відносна частка займенників у тексті
flesch_reading_ease	Індекс читабельності Флеша
gunning_fog_index	Індекс складності (Gunning Fog)

Результати дослідження. Отримані результати показують, що трансформерні моделі демонструють вищу здатність до узагальнення в умовах невідомого генератора порівняно зі стилометричними ансамблями.

На рис. 2 представлено агреговані показники метрики Macro-F1 для всіх моделей у кожному зі сценаріїв LOGO.

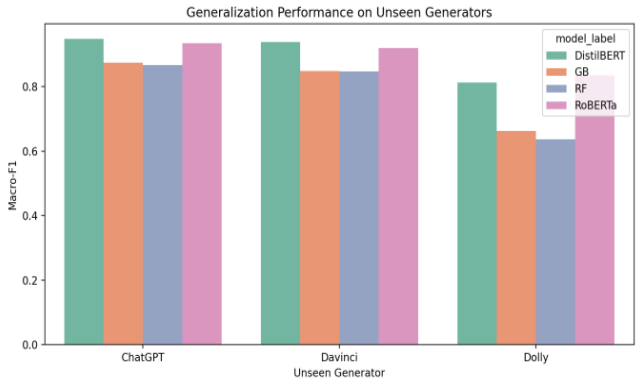

Рис. 2. Загальна здатність моделей до узагальнення у трьох сценаріях

У сценаріях, де тестовими генераторами виступали моделі від OpenAI (ChatGPT та Davinci), DistilBERT та RoBERTa впевнено долають поріг метрики Macro-F1 у 0.85–0.95. Водночас класичні моделі демонструють відчутну деградацію. Найбільший розрив спостерігається у сценарії logo_dolly, де тестовою моделлю виступає відкрита архітектура. Зіткнувшись із принципово новим патерном генерації, стилометричні моделі втрачають здатність до ефективного розрізнення класів: Macro-F1 для Random Forest падає до значень близько 0.65, що свідчить про їхню неспроможність переносити знання за межі навчального розподілу.

Порівняння стабільності детекторів у розрізі доменів (неформальний Reddit та структурована Wikipedia) підкреслює різну чутливість архітектур до інформаційного шуму та формальності. На рис. 3 візуалізовано показники F1 для класу Machine, а в табл. 3 наведено детальні значення Accuracy, Precision та Recall.

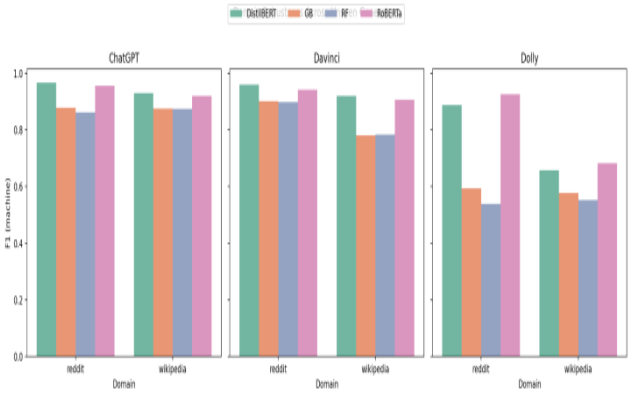

Рис. 3. Порівняння метрики F1 за доменами для різних генераторів

Результати експерименту свідчать, що в умовах генераторного зсуву моделі в більшості випадків демонструють вищі показники на домені Reddit порівняно з доменом Wikipedia. Трансформерні детектори (DistilBERT, RoBERTa) характеризуються високою робастністю, що проявляється у стабільно високих значеннях метрики F1 у домені Reddit. Натомість на енциклопедичних текстах для стилометричного ансамблю Gradient Boosting у сценарії logo_dolly спостерігається суттєве зниження якості детекції, що зумовлено погіршенням здатності моделі виявляти згенеровані тексти.

Таблиця 3 – Показники якості моделей за доменами

Сценарій	Домен	Модель	Accuracy	Precision	Recall
ChatGPT	Reddit	DistilBERT	0.9660	0.9408	0.9946
ChatGPT	Wiki	DistilBERT	0.9276	0.9052	0.9556
ChatGPT	Reddit	GB	0.8775	0.8763	0.8791
ChatGPT	Wiki	GB	0.8707	0.8419	0.9128
ChatGPT	Reddit	RF	0.8634	0.8689	0.8560
ChatGPT	Wiki	RF	0.8699	0.8456	0.9052
ChatGPT	Reddit	RoBERTa	0.9533	0.9157	1.0000
ChatGPT	Wiki	RoBERTa	0.9132	0.8589	0.9909
Davinci	Reddit	DistilBERT	0.9580	0.9263	0.9957
Davinci	Wiki	DistilBERT	0.9154	0.8706	0.9776
Davinci	Reddit	GB	0.8990	0.8788	0.9258
Davinci	Wiki	GB	0.7935	0.8331	0.7342
Davinci	Reddit	RF	0.8961	0.8750	0.9244
Davinci	Wiki	RF	0.7958	0.8317	0.7418
Davinci	Reddit	RoBERTa	0.9390	0.8927	0.9986
Davinci	Wiki	RoBERTa	0.8982	0.8382	0.9882
Dolly	Reddit	DistilBERT	0.8929	0.9333	0.8469
Dolly	Wiki	DistilBERT	0.7297	0.8230	0.5477
Dolly	Reddit	GB	0.6820	0.8204	0.4662
Dolly	Wiki	GB	0.6760	0.7579	0.4657
Dolly	Reddit	RF	0.6529	0.8017	0.4061
Dolly	Wiki	RF	0.6657	0.7549	0.4367
Dolly	Reddit	RoBERTa	0.9258	0.9254	0.9269
Dolly	Wiki	RoBERTa	0.7356	0.7925	0.6024

Перевірка залежності якості детекції від обсягу контексту виявила відмінності в поведінці моделей залежно від довжини тексту. Для цього тестову вибірку було поділено на кошики за довжиною (B2: 101–200 слів, B3: 201–400 слів, B4: понад 400 слів). Загальну динаміку метрик для всіх досліджуваних моделей відображено на рис. 4, тоді як у табл. 4 наведено деталізовані числові результати для двох репрезентативних моделей, а саме, DistilBERT і Random Forest, які представляють відповідно трансформерний і класичний стилометричний підходи.

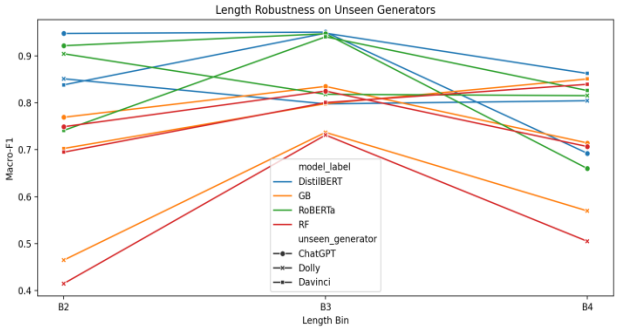

Рис. 4. Зміна метрики Macro-F1 залежно від довжини тексту (Bin Analysis)

Таблиця 4 – Результати оцінювання моделей у розрізі довжини тексту

Група	Модель	Precision	Recall	F1
B2	DistilBERT	0.870±0.017	0.993±0.006	0.927±0.012
	RF	0.667±0.007	0.603±0.016	0.633±0.012
B3	DistilBERT	0.964±0.008	0.978±0.006	0.971±0.003
	RF	0.892±0.001	0.899±0.003	0.895±0.002
B4	DistilBERT	0.294±0.031	0.829±0.029	0.433±0.03
	RF	0.325±0.008	0.743±0.000	0.452±0.008

На коротких текстах (B2) класичні моделі (RF) демонструють низькі значення метрики Recall (0.603). Із збільшенням обсягу тексту (B3) якість класифікації покращується, що пов'язано з накопиченням статистичних характеристик, необхідних для оцінювання лексичного різноманіття. Водночас для довгих текстів (B4) спостерігається суттєве зниження метрики Precision як для класичних, так і для трансформерних моделей (до 0.294 для DistilBERT та 0.325 для RF). Це свідчить про зростання кількості хибно позитивних класифікацій, коли тексти, написані людиною, помилково визначаються як згенеровані. Для трансформерних моделей така поведінка може бути пов'язана з обмеженням максимальної довжини контексту (512 токенів), що призводить до втрати частини релевантної інформації під час обробки довгих текстів. У результаті детектори можуть некоректно інтерпретувати структурні характеристики довгих текстів, що негативно впливає на якість класифікації.

Окремим важливим аспектом є аналіз хибно негативних помилок. Аналіз абсолютної кількості хибно негативних помилок (False Negatives), коли згенерований текст пропускається класифікатором і позначається як людський, є ключовим для розуміння деградації стилометричних підходів. На рис. 5 наведено розподіл таких помилок.

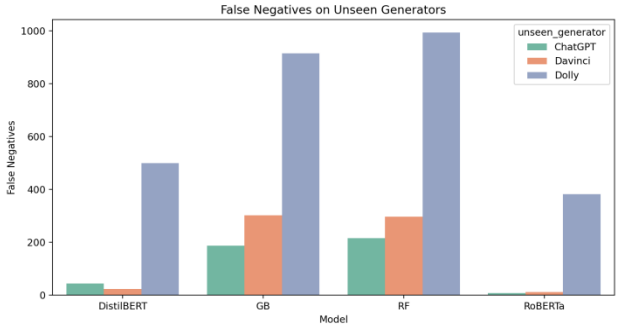

Рис. 5. Абсолютна кількість хибнонегативних помилок у різних моделях

Графік наочно підтверджує, що головною причиною деградації класичних алгоритмів є суттєве зростання кількості пропусків машинного тексту. У сценарії з моделлю Dolly стилометричні ансамблі (GB, RF) генерують від 500 до майже 1000 хибнонегативних відповідей, тоді як трансформери (DistilBERT, RoBERTa) мінімізують цей показник. Зіткнувшись з новою LLM, класичний детектор не знаходить звичних статистичних відхилень і класифікує текст як людський.

Для пояснення механіки перенавчання (overfitting) стилометричних ансамблів ми проаналізували зсув важливості ознак (Feature Importance Shift) між сценаріями LOGO. Візуалізація цього процесу представлена на рис. 6 та в табл. 5.

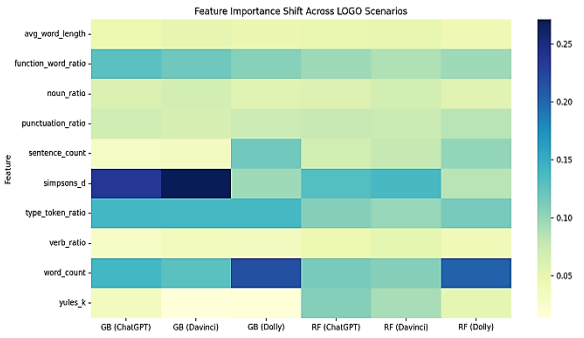

Рис. 6. Теплова карта зсуву важливості стилометричних ознак між сценаріями

Таблиця 5 – Топ-3 найважливіших ознак для алгоритму Gradient Boosting за сценаріями

Сценарій	Топ-1	Топ-2	Топ-3
ChatGPT	simpsons_d (0.237)	word_count (0.141)	type_token_ratio (0.141)
Davinci	simpsons_d (0.271)	type_token_ratio (0.140)	word_count (0.129)
Dolly	word_count (0.218)	type_token_ratio (0.140)	sentence_count (0.117)

У сценаріях, де алгоритми навчалися і тестувалися в межах екосистеми OpenAI (ChatGPT, Davinci), домінуючою ознакою для прийняття рішення був індекс лексичного різноманіття Сімпсона (simpsons_d), вагове значення якого сягало 0.271. Це свідчить про те, що моделі відображають специфічні закономірності словникового розподілу, характерні для текстів, згенерованих моделями типу GPT. Водночас у сценарії, де невідомим генератором виступає Dolly, значущість цієї ознаки суттєво знижується (до 0.097), що вказує на втрату її дискримінаційної здатності.

За відсутності стабільних індикаторів детектори переорієнтовуються на більш загальні поверхневі характеристики, зокрема кількість слів (word_count = 0.218) та речень. Оскільки такі ознаки не є надійними маркерами штучного походження тексту, це призводить до зниження якості класифікації.

Висновки. У роботі реалізовано відтворюваний експеримент для оцінювання змін якості детекторів в умовах появи невідомого генератора. Дослідження базується на застосуванні протоколу LOGO та охоплює дві модельні парадигми, а саме класичні стилометричні підходи і трансформерні encoder-детектори.

Результати дослідження підтверджують вищу здатність трансформерних моделей до узагальнення в умовах появи невідомого генератора. Водночас стилометричні моделі демонструють суттєве зниження ефективності, що пов'язано з їхньою залежністю від поверхневих характеристик тексту, які не забезпечують стабільності при зміні генератора.

Ключовим фактором деградації класичних підходів є зростання кількості хибно негативних помилок, коли згенерований текст класифікується як людський. Такий тип помилок безпосередньо знижує практичну придатність детекторів

у реальних умовах застосування. Ризики їх деградації детекторів і значущість помилок обох типів також відзначаються у дослідженнях, присвячених протидії нейро-фейкам [12].

Дослідження має низку обмежень, пов'язаних із вибором генераторів і доменів, а також із використанням обмеження довжини контексту для трансформерних моделей. Крім того, не розглядалися сценарії часткового або змішаного авторства в межах одного документа.

Подальші дослідження доцільно спрямувати на розширення множини генераторів, залучення багатомовних даних та аналіз гібридних підходів до детекції. Як перспективний напрям можна розглядати підходи, що не потребують навчання з учителем і базуються на аналізі властивостей ймовірнісного простору мовних моделей, зокрема DetectGPT[13].

Декларація про використання генеративного штучного інтелекту. Під час підготовки цієї роботи автори використовували ChatGPT та Gemini для перевірки граматики та орфографії, перефразування та переформулювання тексту. Після використання цих інструментів/сервісів автори перевірили та відредагували вміст за необхідності та несуть повну відповідальність за зміст публікації.

References

1. Wang Y., Mansurov J., Ivanov P., et al. M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection. *EACL 2024*. 2024. Available at: https://aclanthology.org/2024.eacl-long.83.pdf (accessed 05.03.2026).
2. Schuster T., Schuster R., Shah D. J., Barzilay R. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*. 2020, vol. 46, no. 2, pp. 499–510. DOI: 10.1162/coli_a_00380.
3. Pu X., Zhang J., Han X., Tsvetkov Y., He T. On the Zero-Shot Generalization of Machine-Generated Text Detectors. *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 4799–4808. DOI: 10.18653/v1/2023.findings-emnlp.318.
4. Databricks. *Databricks Dolly (Dolly-v2) — official repository/documentation.* Available at: https://github.com/databrickslabs/dolly (accessed 05.03.2026).
5. scikit-learn developers. *GroupShuffleSplit — scikit-learn documentation*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupShuffleSplit.htm (accessed 05.03.2026)
6. imbalanced-learn developers. *Under-sampling — imbalanced-learn user guide (v0.14.1)*. Available at: https://imbalanced-learn.org/stable/under_sampling.html (accessed 05.03.2026).
7. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv, Computer Science*. 2019. DOI: 10.48550/arXiv.1910.01108. Available at: https://arxiv.org/abs/1910.01108 (accessed 05.03.2026).
8. Liu Y., Ott M., Goyal N., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv, Computer Science*. 2019. DOI: 10.48550/arXiv.1907.11692. Available at: https://arxiv.org/abs/1907.11692 (accessed 05.03.2026).
9. Breiman L. Random Forests. *Machine Learning*. 2001, vol. 45, pp. 5–32. DOI: 10.1023/A:1010933404324. Available at: https://link.springer.com/article/10.1023/A:1010933404324 (accessed 05.03.2026).
10. Friedman J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 2001, vol. 29, no. 5, pp. 1189–1232. DOI: 10.1214/aos/1013203451. Available at: https://www.cse.cuhk.edu.hk/irwin.king/_media/presentations/2001_greedy_function_approximation_a_gradient_boosting_machine.pdf (accessed 05.03.2026).
11. Zheng A., Casari A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, O'Reilly Publ., 2018. Available at: https://dl.acm.org/doi/abs/10.5555/3239815 (accessed 05.03.2026).
12. Zellers R., Holtzman A., Rashkin H., et al. Defending Against Neural Fake News. *arXiv, Computer Science*. 2019. DOI: 10.48550/arXiv.1905.12616. Available at: https://arxiv.org/abs/1905.12616 (accessed 05.03.2026).
13. Mitchell E., Lee Y., Khazatsky A., Manning C. D., Finn C. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *Proceedings of Machine Learning Research*. 2023, vol. 202. Available at: https://proceedings.mlr.press/v202/mitchell23a.html (accessed 05.03.2026).

Надійшла (received) 10.03.2026
Прийнята (accepted) 30.03.2026
Оприлюднена (published) 20.05.2026

UDC 004.89:519.76

T. V. PETRYSHAK, Postgraduate Student at the Department of Information Systems and Networks, Lviv Polytechnic National University, Lviv, Ukraine; e-mail: taras.v.petryshak@lpnu.ua; ORCID: https://orcid.org/0009-0006-6296-3867
V. A. VYSOTSKA, Doctor of Technical Sciences, Professor at the Department of Information Systems and Networks, Lviv Polytechnic National University, Lviv, Ukraine; e-mail: Victoria.A.Vysotska@lpnu.ua; ORCID: https://orcid.org/0000-0001-6417-3689

EVALUATING THE GENERALIZATION ABILITY OF AI-GENERATED TEXT DETECTORS TO UNSEEN GENERATORS

Many AI-generated text detectors demonstrate high performance on datasets constructed within typical evaluation protocols. In particular, classical models based on stylometric features, such as text length, punctuation patterns, and aggregated formality indicators, can effectively capture statistical regularities of machine generation. However, their performance decreases substantially when texts produced by previously unseen generators are encountered. Under such conditions, feature distributions shift, which leads to a decline in classification quality, primarily due to an increase in false negative errors. This paper investigates the generalization ability of detection models under conditions involving an unseen generator. The study compares classical stylometric models and transformer-based approaches using the LOGO (Leave-One-Generator-Out) evaluation protocol. The task is formulated as binary text classification across two domains, Reddit and Wikipedia, and involves three generators, namely ChatGPT, Davinci, and Dolly. The classical models include Random Forest and Gradient Boosting, whereas the transformer-based approaches are represented by DistilBERT and RoBERTa. Model performance is evaluated using Accuracy, Precision, Recall, F1, and Macro-F1, with the final results averaged across multiple random initializations. The results show that transformer-based models demonstrate a higher ability to generalize to texts produced by unseen generators. In contrast, stylometric approaches exhibit a substantial degradation in performance, particularly depending on the domain and text length. Error analysis indicates that the main factor behind this decline is the increase in false negative errors. An additional analysis of feature importance shows that classical models rely heavily on surface-level textual characteristics, which do not ensure stable generalization across different generators. Therefore, the findings highlight the importance of evaluating AI-generated text detectors under the LOGO protocol to ensure robust performance in the presence of new language models.

Keywords: AI-generated text detection, generalization ability, unseen generator, stylometric features, transformer-based models, text classification.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Петришак Тарас Володимирович / Petryshak Taras Volodymyrovych
Автор 2 / Author 2: Висоцька Вікторія Анатоліївна / Vysotska Viktoriia Anatoliivna