ДОДАТОК А

Слайди презентації

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Дослідження мовних моделей для виявлення текстів, згенерованих ШІ

Ткаченко Олександра Олексіївна, гр. ІПЗм-23-3
Науковий керівник: проф. Смеляков Кирило Сергійович

13 червня 2025

Дослідження

Мовні моделі стали основою сучасних технологій обробки природної мови. Особливої популярності набули великі мовні моделі (LLM), такі як ChatGPT або Mistral 7B, здатні генерувати тексти, що майже не відрізняються від людських.

Швидке поширення LLM супроводжується низкою викликів, зокрема зростає ризик використання штучно згенерованого контенту в академічному та публічному середовищі, що ставить під загрозу доброчесність, достовірність інформації та дотримання авторських прав.

За даними Liang, W., Zhang та ін. "Mapping the Increasing Use of LLMs in Scientific Papers" (2024), які проаналізували 950 965 наукових публікацій за 2020–2024 роки, до 17,5 % текстів у галузі комп'ютерних наук містять ознаки використання великих мовних моделей.

Основні підходи до виявлення текстів згенерованих ШІ:

- **Статистичні**. Аналіз перплексії, ентропії, n-грам. Не потребують навчання, але малоефективні для сучасних LLM.
- **Класифікатори**. Навчаються розрізняти людські й AI-тексти. Мовні моделі (BERT, DeBERTa) використовуються для витягування ознак.
- **Водяні знаки**. Вбудовуються під час генерації тексту. Працюють лише за підтримки з боку генератора.

У цьому дослідженні ми зосередимося на другому підході - класифікаторах на основі мовних моделей.

Об'єктом дослідження є мовні моделі, їх ефективність та можливості застосування для розпізнавання текстів, створених штучним інтелектом.

2

Огляд літератури

- *"A Survey of Large Language Models"* (Zhao et al., 2024): Класифікація мовних моделей, аналіз їх можливостей та обмежень.
- *"Mapping the Increasing Use of LLMs in Scientific Papers"* (Liang et al., 2024): Дослідження поширення LLM у наукових публікаціях.
- *"On the Possibilities of AI-Generated Text Detection"* (Chakraborty et al., 2024): Огляд сучасних методів детекції AI-текстів (статистичних, класифікаційних, комбінованих).
- *"How to Detect AI-Generated Texts?"* (Nguyen et al., 2023): Методи виявлення AI-текстів через ознаки подібності та векторні представлення.
- *"Attention Is All You Need"* (Vaswani et al., 2017): Запропонована архітектура Transformer, що стала основою сучасних моделей NLP.

Постановка задачі

Проблема:
Штучно згенеровані тексти дедалі частіше використовуються в науці, освіті та медіа. Виникає потреба у надійних інструментах, здатних розпізнавати такі тексти. Водночас ефективність мовних моделей як засобу ідентифікації залишається недостатньо дослідженою.

Мета дослідження:
Проаналізувати ефективність використання мовних моделей для задачі розпізнавання текстів, згенерованих штучним інтелектом, та встановити найбільш придатні для практичного використання.

Очікувані результати:
Сформований набір критеріїв та ранжування мовних моделей за ефективністю, ресурсною доцільністю та практичністю застосування.

Методологія

- **Методи дослідження:**
  – Класифікація мовних моделей за типами (SLM, NLM, PLM, LLM)
  – Побудова векторної моделі з оціночними критеріями
  – Застосування згортки для визначення оптимальних моделей
  – Експериментальне порівняння на основі реальних даних
- **Використані технології та інструменти:**
  – Python, бібліотеки Hugging Face Transformers
  – Google Colab з GPU (A100)
  – Дата-сет ChatGPT-Research-Abstracts (20 000 текстів)
  – Метрики: Accuracy, Precision, Recall, F1-score

Теоретична частина експеременту

З метою виявлення найефективніших моделей для розпізнавання AI-згенерованих текстів було здійснено порівняння мовних моделей, що представляють чотири етапи їхнього розвитку:

- **SLM -** статистичні мовні моделі
- **NLM** - нейронні моделі,
- **PLM** - попередньо навчені моделі,
- **LLM** - великі мовні моделі.

Із кожної групи було обрано найбільш перспективні представники. Для об'єктивного порівняння моделей використано 5 критеріїв:

- Кількість параметрів
- Максимальний контекст
- Обсяг споживаної пам'яті
- Розмір словника
- Призначення для класифікації тексту

Модель	Кількість параметрів (у мільйонах)	Контекст (токени)	Необхідна пам'ять (GB)	Розмір словника (токени)	Призначення для класифікації тексту
Modified Kneser-Ney	Немає параметрів	3	1	20 000	Ні
FastText	Необмежена	10	4	Необмежена	Так
ELMo	30 (Маленька модель)	512	8	50 000	Ні
BERT	110 (Середня модель)	512	12	30 000	Так
RoBERTa	125 (Середня модель)	512	16	50 000	Так
DeBERTa-v3	184 (Середня модель)	512	16	128 000	Так
Phi-2	2700 (Велика модель)	2048	5.6	51 200	Ні
Orion-14B	14000 (Велика модель)	4096	29	84 608	Ні
Mistral 7B	7000 (Велика модель)	32K	14.4	32 000	Ні
OpenChat-3.5	7000 (Велика модель)	8192	14.4	32 002	Ні

Результати теоретичної частини

Кроки обробки:

1. Приведення шкал до принципу оптимальності
Критерії, де кращим є менше значення (наприклад, споживання пам'яті), трансформуються до зворотної шкали.

2. Фільтрація за принципом Парето
Альтернативи, що гірші за інші за всіма критеріями, виключаються.

3. Нормалізація значень
Для кожного критерію *j* і моделі *i*:

$$f_{ij}^{\text{norm}} = \frac{f_{ij} - f_j^{\min}}{f_j^{\max} - f_j^{\min}}$$

4. Обчислення нормуючого множника
Для кожного критерію *j* і моделі *i*:

$$a_j = \frac{1}{\sum_{i=1}^{m} a_{ij}}$$

5. Лінійна згортка з вагами

$$Z^* = \max_{i=1,\ldots,m} \sum_{j=1}^{n} a_j b_j a_{ij}$$

де b_j - ваговий коефіцієнт критерію,
a_{ij} - нормалізоване значення.

Модель	Результат згортки
Modified Kneser-Ney	0.0144543
FastText	0.2024076
BERT	0.0897989
DeBERTa-v3	0.1346456
Phi-2	0.0686771
Orion-14B	0.1142685
Mistral 7B	0.2557635
OpenChat-3.5	0.1199842

Отже для подальшого дослідження будуть використані чотири моделі що показали найвищий результат, а саме: FastText, Mistral 7B, DeBERTa-v3, OpenChat-3.5.

7

Практична частина експеременту

Для експерименту обрано моделі FastText, DeBERTa-v3, Mistral 7B та OpenChat-3.5 — найбільш перспективні за результатами теоретичного аналізу.

Дані: використано датасет ChatGPT-Research-Abstracts (Hugging Face) — 20 000 англомовних пар "людина–ШІ", згенерованих GPT-3.5.

Попередня обробка: зниження регістру, видалення шуму та спецсимволів, очищення stop-words.

Інструменти: Python 3.x, бібліотеки: transformers, datasets, peft, pandas, nltk, re.

Середовище: Google Colab (GPU Nvidia A100).

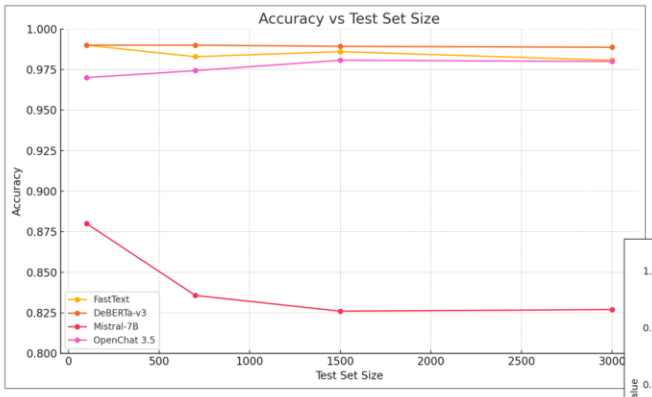
Схема експерименту:

- поділ даних на train (15k), validation (2k), test (3k);
- донавчання моделей на тренувальній вибірці;
- налаштування на validation, оцінка на test;
- метрики: Accuracy, Precision, Recall.

Модель	Середній час тестування (3000 текстів)	Мінімальний обсяг відеопам'яті
FastText	<1с	0 гб
DeBERTa-v3	5 хв 40 с	4 гб
Mistral-7B	38 хв	18 гб
OpenChat 3.5	29 хв	16 гб

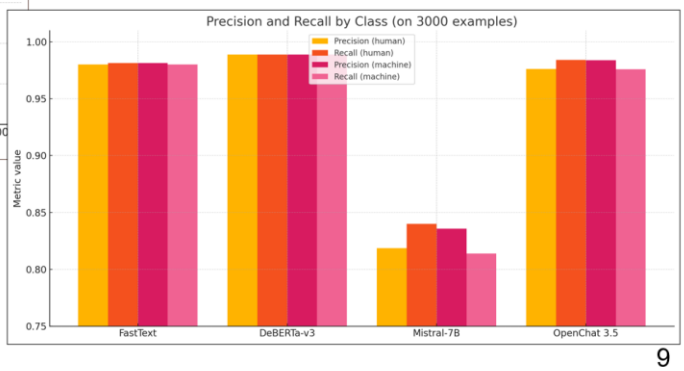
8

Результати практичної частини



$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

9

Аналіз отриманих результатів

Найвищу точність класифікації показала **DeBERTa-v3**, яка забезпечує добрий баланс між якістю та обчислювальними витратами. **FastText** досягла порівнянної ефективності при мінімальних ресурсах і дуже високій швидкості без використання GPU.

OpenChat 3.5 показала гарні, але нижчі за точністю результати, і потребує понад 16 ГБ пам'яті. **Mistral-7B** виявилась найменш придатною через нижчу точність і значне ресурсне навантаження.

Завдання виявлення AI-текстів не вимагає глибокого контекстуального аналізу, тому великі моделі не демонструють переваг над компактнішими рішеннями.

Рекомендації:

- **FastText** — для швидкої класифікації за умов обмежених ресурсів;
- **DeBERTa-v3** — коли пріоритетом є точність;
- **LLM** — доцільні для генерації, а не класифікації текстів.

Перспективним напрямом подальших досліджень є поєднання мовних моделей із альтернативними підходами — стилометрією, синтаксичним аналізом або виявленням структурних патернів.

10

Публікація результатів

MIT@AIS-2025

Research On Language Models For Detecting Texts Generated By Artificial Intelligence

Oleksandra Tkachenko[1] and Kyrylo Smelyakov[1]

[1] Kharkiv National University of Radio Electronics, Nauky ave., 14, Kharkiv, 61166, Ukraine

Abstract. This paper explores current approaches for detecting texts generated by artificial intelligence, focusing on three main strategies: statistical methods, classification models, and watermarking techniques. It introduces a classification of language models into four stages: statistical, neural, pretrained, and large, analyzing their evolution and suitability for text detection tasks. The study highlights key challenges, including the increasing similarity of AI-generated texts to human writing, the need for high-quality labeled data, and the computational demands of modern detection systems. A theoretical evaluation of models based on criteria such as parameter count, context window, and memory requirements identified FastText, Mistral 7B, DeBERTa-v3, and OpenChat-3.5 as top candidates. These models were further tested in a practical experiment using a balanced dataset of human- and AI-written texts. Results showed that DeBERTa-v3 achieved the highest accuracy, while FastText offered strong performance with minimal resources. The findings suggest that medium-scale models are more effective for AI-text detection than large language models.

Keywords: Classification, Language Models, Transformers.

1 Introduction

Language modeling is one of the key areas in the development of artificial intelligence, evolving from statistical models to large transformer-based architectures. Large language models, such as ChatGPT, are capable of generating coherent and grammatically correct texts. They are widely used in machine translation, classification, and dialogue systems [1].

Despite their high effectiveness, LLMs pose several risks, from spreading misinformation to violating copyright and undermining academic integrity. An analysis of over 950,000 scientific papers has shown a rapid increase in the use of LLMs, especially in computer science [2], which highlights the growing need for reliable tools to detect machine-generated content.

This paper explores the main approaches to detecting AI-generated texts, presents a classification of language models, and outlines the current limitations and future development prospects of these methods.

11

Дякую за увагу!

12

ДОДАТОК Б

Апробація результатів роботи

MIT@AIS-2025

19 - 22 may
Kharkiv-Yaremche 2025

Certificate

Oleksandra Tkachenko

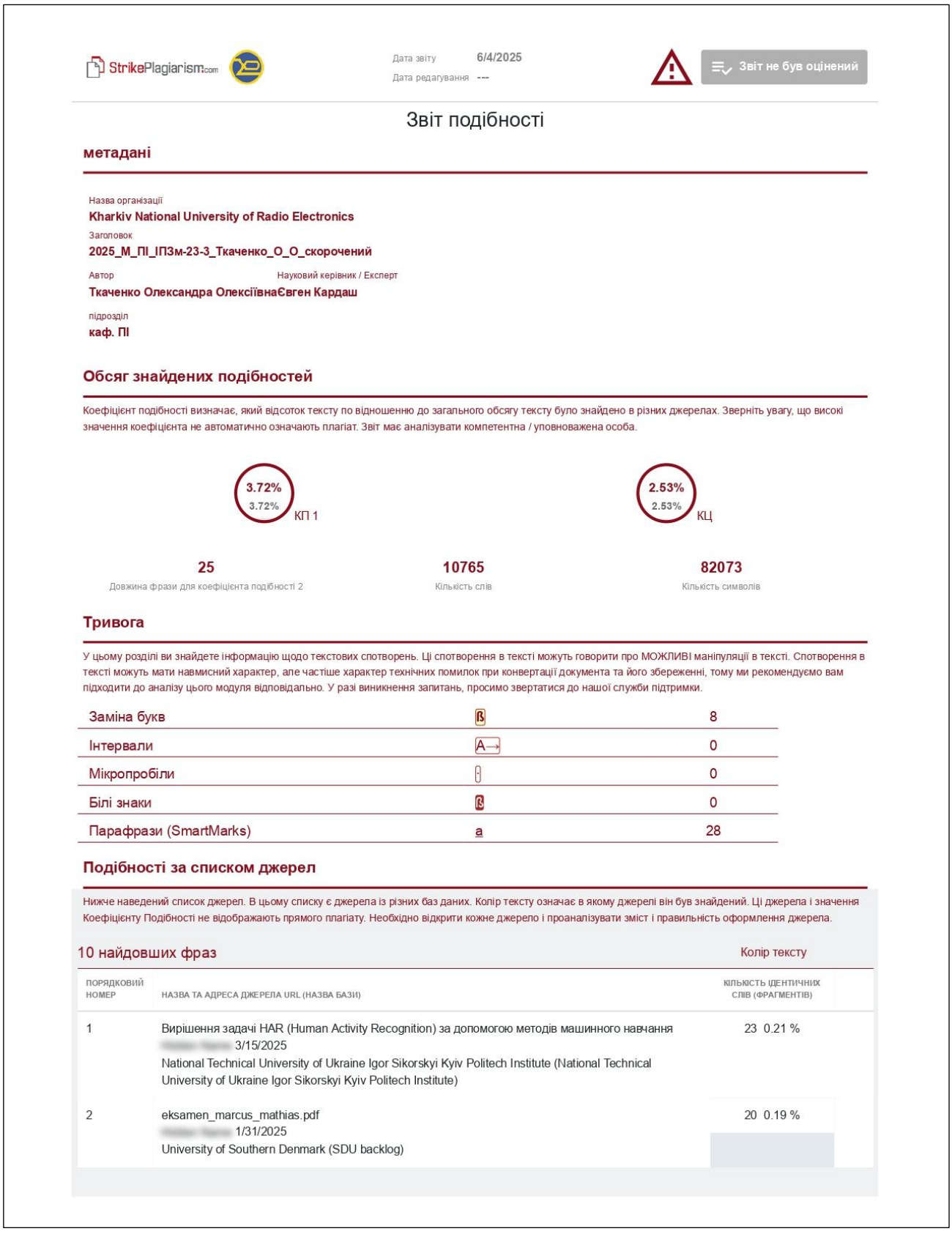
particapated in
1st International Scientific and Practical Conference
"Modern Information Technologies and
Artificial Intelligence Systems"
MIT@AIS-2025

Chair of the conference *Yuri Romanenkov*

Kharkiv 2025

ДОДАТОК В

Результат перевірки на академічний плагіат



StrikePlagiarism.com

Дата звіту 6/4/2025
Дата редагування ---

Звіт не був оцінений

Звіт подібності

метадані

Назва організації
Kharkiv National University of Radio Electronics
Заголовок
2025_М_ПІ_ІПЗм-23-3_Ткаченко_О_О_скорочений
Автор
Ткаченко Олександра ОлексіївнаЄвген Кардаш
Науковий керівник / Експерт
підрозділ
каф. ПІ

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

3.72% 3.72% КП 1

2.53% 2.53% КЦ

25	10765	82073
Довжина фрази для коефіцієнта подібності 2	Кількість слів	Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв	ß	8
Інтервали	A⟶	0
Мікропробіли	·	0
Білі знаки	ß	0
Парафрази (SmartMarks)	a	28

Подібності за списком джерел

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз — Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	Вирішення задачі HAR (Human Activity Recognition) за допомогою методів машинного навчання [redacted] 3/15/2025 National Technical University of Ukraine Igor Sikorskyi Kyiv Politech Institute (National Technical University of Ukraine Igor Sikorskyi Kyiv Politech Institute)	23 0.21 %
2	eksamen_marcus_mathias.pdf [redacted] 1/31/2025 University of Southern Denmark (SDU backlog)	20 0.19 %

ДОДАТОК Г

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008: 2015

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-23-3 (група)

Олександра ТКАЧЕНКО
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
	7.3 Нумерація сторінок звіту	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	

Експерт _____ (підпис) Вадим НЕЧВОЛОД (прізвище, ініціали)

05.06.2025