

Міністерство Освіти і Науки України

Національний університет Києво-Могилянська академія

Кафедра інформаційних технологій, Факультет інформатики



Дослідження методів розпізнавання українських текстів, згенерованих великими мовними моделями (Large Language Models) / Exploring Methods for Detecting Ukrainian Texts Generated by Large Language Models

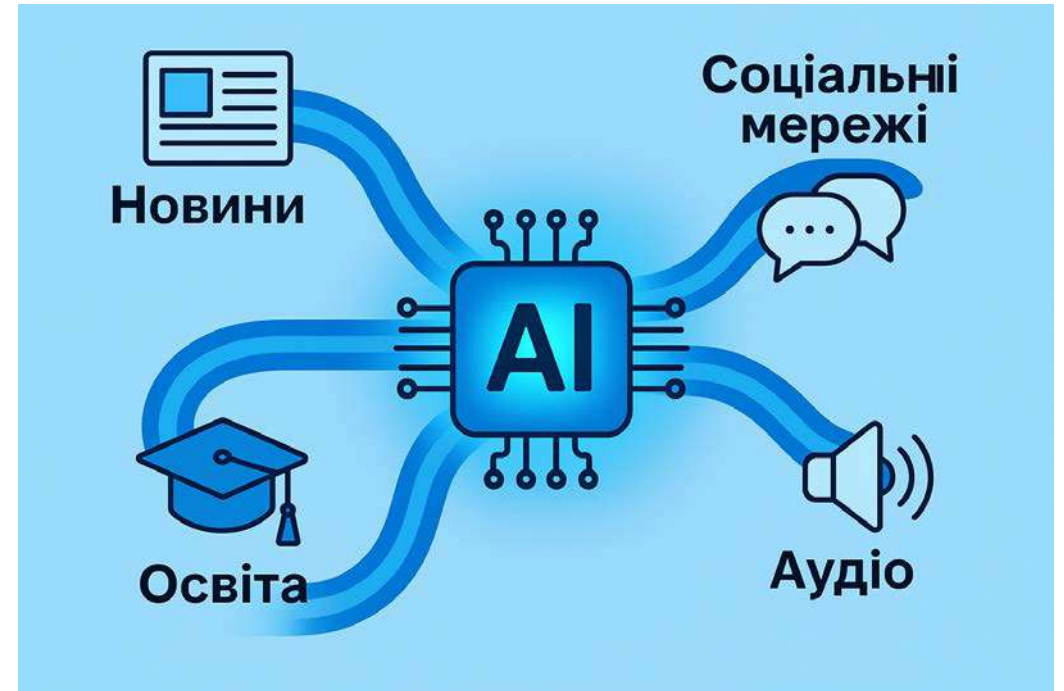
Гоголь Антон

Науковий керівник: Ст. Викладач Кундік К.В.

2025

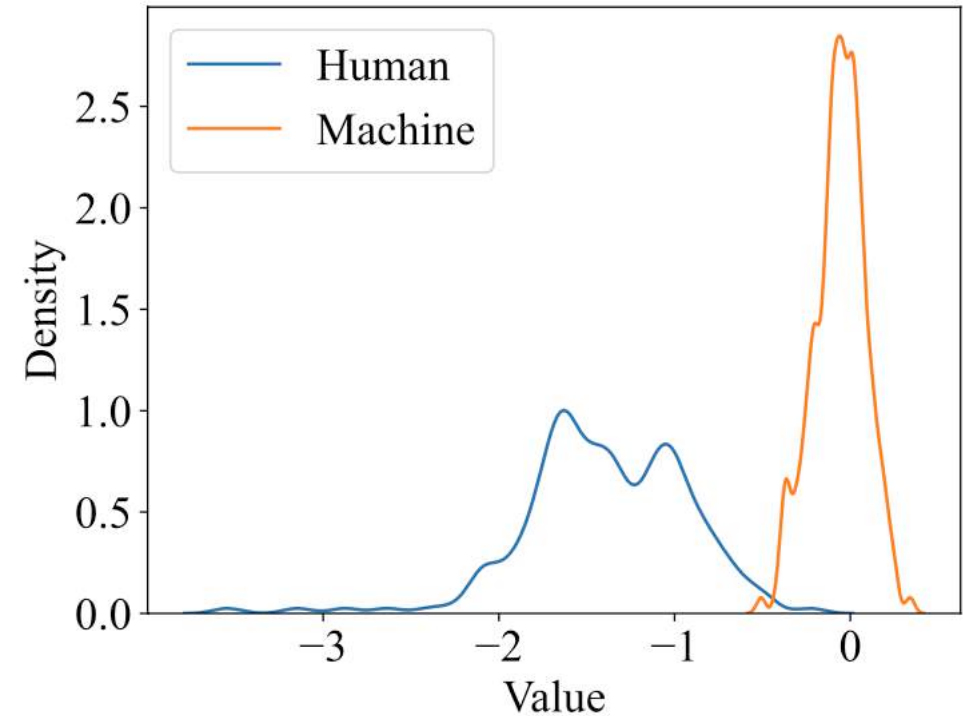
Мета роботи

- Дослідити існуючі методи виявлення тексту, згенерованого мовними моделями.
- Імплементувати та адаптувати ці методи для української мови.
- Протестувати ці методи на україномовному наборі даних.



А як визначати де ШІ?

- Текст згенерований ШІ – детерміністичний.
- Текст згенерований людиною – непередбачуваний.
- Цим можна скористатися.

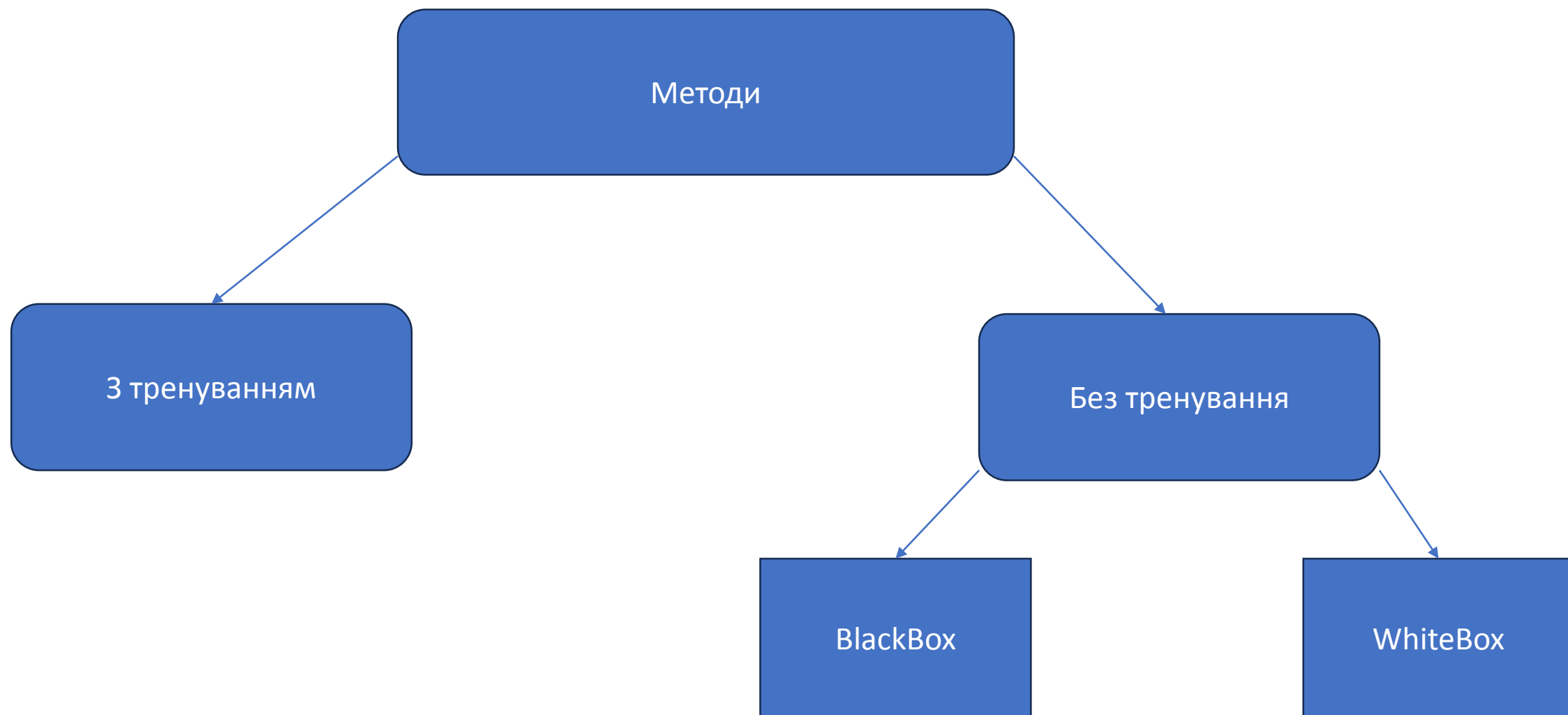


Вісь X – лог. імовірність генерації токена, Y – густина розподілу

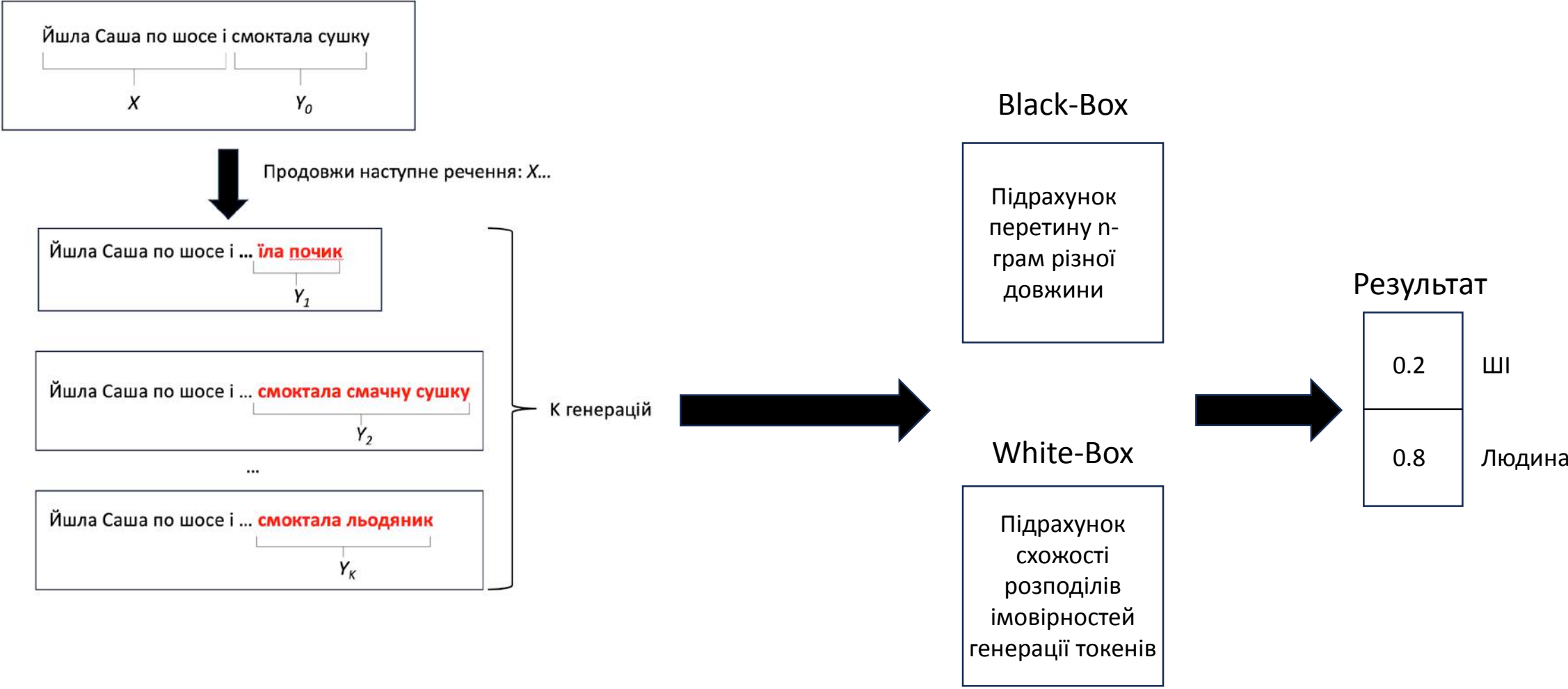
Проблеми українського NLP

- NLP – natural language processing.
- Українська не є такою розповсюженою.
- Якість мовних моделей є гіршою.

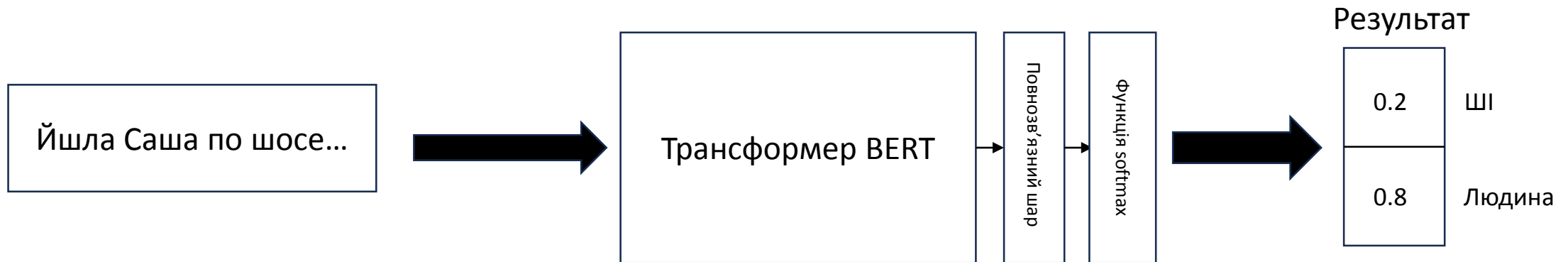
Досліджені методи



DNA-GPT



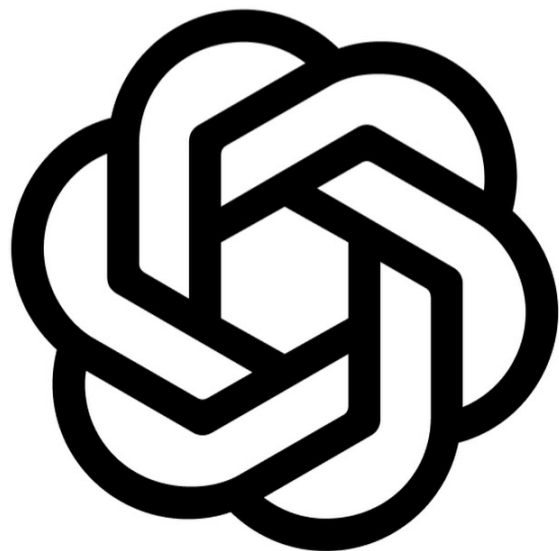
Класифікатор на базі BERT



- BERT (Bidirectional Encoder Representations from Transformers) — це велика мовна модель, запропонована Google у 2018 році.

Використані моделі

- Для генерації даних було використано моделі gpt-4o-mini та Llama3:70B. Для експерименту додатково була використана модель openai davinci-002
- Кожна модель використовувалася для створення трьох окремих наборів даних з різними значеннями параметра температури, а саме 0.1, 0.5 і 0.8



Створення набору даних

Текст	Короткий заголовок	Клас
З нагоди підписання Дейтонських угод 11 років тому Сербія, Боснія-Герцеговина та Чорногорія долучилися до програми НАТО Партнерство заради миру. Генеральний секретар Яап де Хооп Схефер заявив, що Боснія нарешті взяла курс на євроатлантичну інтеграцію...	Три балканські країни - Боснія-Герцеговина, Чорногорія і Сербія - офіційно підписали угоду, що передуює їхньому приєднанню до НАТО.	Human
LLM.input(Prompt = "Напиши новину з {length} слів для наступного заголовку: {заголовок}", Temperature =0.1)	Три балканські країни - Боснія-Герцеговина, Чорногорія і Сербія - офіційно підписали угоду, що передуює їхньому приєднанню до НАТО.	AI

Результати Black-box DNA-GPT

Набір даних	Модель для генерації продовження	FPR	TPR
Human + gpt-4o 0.1 temp	gpt-4o-mini	<1%	48.6%
Human + gpt-4o 0.5 temp	gpt-4o-mini	<1%	41.1%
Human + gpt-4o 0.8 temp	gpt-4o-mini	<1%	31.5%
Human + llama 0.1 temp	llama	<1%	27%
Human + llama 0.5 temp	llama	<1%	25.1%
Human + llama 0.8 temp	llama	<1%	20.8%
Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	gpt-4o-mini	<1%	<1%
Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	llama	<1%	<1%
Human + gpt-4 0.7 temp ENGLISH	gpt-4	<1%	32.6%

Результати White-box DNA-GPT

Набір даних	Модель для генерації продовження	FPR	TPR
Human + gpt-4o 0.1 temp	davinci	<1%	10.4%
Human + gpt-4o 0.5 temp	davinci	<1%	10.2%
Human + gpt-4o 0.8 temp	davinci	<1%	9.4%
Human + llama 0.1 temp	davinci	<1%	<1%
Human + llama 0.5 temp	davinci	<1%	<1%
Human + llama 0.8 temp	davinci	<1%	<1%
Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	davinci	<1%	<1%
Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	davinci	<1%	<1%
Human + davinci 0.7 temp ENGLISH	davinci	<1%	99%

Результати класифікатора на базі BERT

Тренувальний набір даних	Тестовий набір даних	FPR	TPR
Human + gpt-4o 0.1 temp	Human + gpt-4o 0.1 temp	<1%	99.3%
Human + gpt-4o 0.5 temp	Human + gpt-4o 0.5 temp	<1%	98.4%
Human + gpt-4o 0.8 temp	Human + gpt-4o 0.8 temp	<1%	98.6%
Human + llama 0.1 temp	Human + llama 0.1 temp	<1%	78.2%
Human + llama 0.5 temp	Human + llama 0.5 temp	<1%	80.1%
Human + llama 0.8 temp	Human + llama 0.8 temp	<1%	77.9%
Human + gpt-4o 0.1, 0.5, 0.8 temp	Human + gpt-4o 0.1, 0.5, 0.8 temp	<1%	90.4%
Human + gpt-4o 0.1, 0.5, 0.8 temp	Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	<1%	31.5%
Human + llama 0.1, 0.5, 0.8 temp	Human + llama 0.1, 0.5, 0.8 temp	<1%	52.1%
Human + llama 0.1, 0.5, 0.8 temp	Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	<1%	22.1%
Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	Human + llama + gpt-4o 0.1, 0.5, 0.8 temp	<1%	77.5%

Висновки

- Незважаючи на проблеми українського NLP – виявлення можливе і були імплементовані декілька алгоритмів, що виявилися дієвими.
- Навіть методи, що не потребують навчання і доступу до ваг моделей є ефективним.
- Дієвість методу дуже сильно залежить від генеративної моделі і температури тексту.
- За більших обчислюваних можливостей і більш широкому наборі даних – метод можна розширити не лише на одну вузьку тему текстів.

