# МЕТОДОЛОГІЯ ВИЯВЛЕННЯ AI-ЗГЕНЕРОВАНОГО ТЕКСТУ УКРАЇНСЬКОЮ МОВОЮ

Ємельяненко А., Просолов В.В.
Харківський національний університет радіоелектроніки, Харків, Україна

З стрімким розвитком генеративного штучного інтелекту дедалі складніше стає відрізнити текст, згенерований штучним інтелектом (AI-generated text), від тексту, написаного людиною. Це створює низку проблем у суспільстві та наукових дослідженнях, зокрема сприяє поширенню фейкових новин, академічного плагіату та дезінформації.

Актуальною є проблема виявлення AI-згенерованого тексту, оскільки сучасні моделі здатні створювати тексти, які практично не відрізняються від написаних людиною.

Особливо це актуально для української мови. У порівнянні з англійською вона має складнішу морфологію, іншу структуру речень і меншу представленість у тренувальних даних мовних моделей. Це означає, що універсальні підходи детекції працюють менш ефективно і потребують адаптації.

**Метою доповіді** є розробка підходу до виявлення AI-згенерованого тексту українською мовою з урахуванням її мовних особливостей.

Сучасна методологія аналізу ефективності детекторів AI-тексту базується на різноманітних підходах до виявлення характерних ознак машинного авторства. Однією з основних стратегій є використання нейронних класифікаторів, які навчаються на корпусах текстів, створених людиною та штучним інтелектом. Наприклад, моделі на основі RoBERTa, донавчені для розрізнення людських і AI-текстів, демонструють здатність виявляти тонкі відмінності у стилі та структурі [1]. Інший підхід полягає у використанні статистичних метрик, таких як ентропія або перплексія. Інструменти на кшталт DetectGPT аналізують розподіл ймовірностей слів у тексті, виявляючи підвищену передбачуваність, притаманну AI-генерації.

Далі виконується аналіз текстів за кількома групами ознак. До них належать особливості словникового складу, структура речень, частота використання певних слів, а також загальна зв’язність тексту. Окремо враховуються морфологічні та синтаксичні характеристики, які мають важливе значення саме для української мови.

Ключовою проблемою у створенні інструкційно-налаштованих моделей для української мови, оптимізованих для редагування тексту, є обмежена доступність відповідних датасетів. Для вирішення цієї проблеми адаптують наявні датасети з української та англійської мов, перетворюючи їх на датасети для виконання інструкцій. Це дозволяє будувати сучасні моделі редагування тексту для української мови, яка здатна виконувати складні завдання редагування: граматичну корекцію, спрощення тексту, забезпечення зв’язності та перефразування [1].

Керовані методи виявлення AI-згенерованого тексту базуються на використанні розмічених датасетів, де кожен текст має мітку, що вказує на його походження - людське чи машинне. Ці підходи розглядають задачу як бінарну класифікацію, навчаючи моделі розрізняти тексти за допомогою різноманітних ознак.

Дослідження у цій сфері демонструють, що керовані методи зазвичай забезпечують високу точність виявлення, особливо коли для навчання доступна велика кількість якісних даних.

Проте, ці методи мають суттєві обмеження: вони схильні до перенавчання, особливо у випадках, коли навчальні дані обмежені за обсягом або специфічні для певної доменної області. Це може призводити до зниження ефективності при застосуванні моделей до текстів, що відрізняються від тих, на яких вони були навчені [2].

Гібридні підходи поєднують традиційні ознаки тексту, такі як кількість слів, багатство словника, індекси читабельності, з машинним навчанням або нейронними моделями.

Такі методи дозволяють враховувати як поверхневі, так і глибші лінгвістичні характеристики тексту, що підвищує загальну точність виявлення.

Часто для цього використовують ансамблеві стратегії, які комбінують результати кількох моделей або різних типів ознак, що дозволяє зменшити ймовірність помилкової класифікації [2].

Однією з ключових проблем для подальшого розвитку методології виявлення AI-згенерованого тексту українською мовою є обмежена доступність якісних датасетів для навчання та тестування моделей, орієнтованих на завдання редагування та аналізу тексту.

У відповідь на цю проблему дослідники адаптують існуючі датасети з української та англійської мов, перетворюючи їх на інструкційно-орієнтовані набори даних, що дозволяє будувати сучасні моделі для редагування тексту українською.

Такі підходи відкривають нові можливості для створення моделей, здатних не лише генерувати, а й аналізувати та розпізнавати AI-згенерований текст, враховуючи специфіку української мови [3].

**Список літератури**

1. A. Saini, A. Chernodub, V. Raheja, and V. Kulkarni, "Spivavtor: An Instruction Tuned Ukrainian Text Editing Model," arXiv:2404.18880, 2024. URL: https://arxiv.org/abs/2404.18880

2. S. Agrahari and S. R. Singh, "Tracing Thought: Using Chain-of-Thought Reasoning to Identify the LLM Behind AI-Generated Text," arXiv:2504.16913, 2025. URL: https://arxiv.org/abs/2504.16913

3. I. David and A. Gervais, "AuthorMist: Evading AI Text Detectors with Reinforcement Learning," arXiv:2503.08716, 2025. URL: https://arxiv.org/abs/2503.08716