УДК 004.8+004.056

Л. М. Куперштейн
В. О. Клименко

МЕТОДИ ТА ЗАСОБИ ВИЯВЛЕННЯ ФЕЙКОВОГО КОНТЕНТУ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ: ОГЛЯД СУЧАСНИХ ПІДХОДІВ

Вінницький національний технічний університет

Анотація

У тезах представлено огляд сучасних методів та засобів виявлення фейкового контенту із застосуванням штучного інтелекту. Запропоновано класифікацію підходів за чотирма модальностями – текст, зображення, відео та аудіо – з виокремленням наскрізного рівня мультимодального аналізу. Розглянуто трансформерні детектори фейкових і згенерованих текстів, методи виявлення дипфейків на основі згорткових та трансформерних нейронних мереж, а також системи виявлення синтезованого мовлення із застосуванням самокерованого навчання. Показано, що сучасні детектори демонструють високу точність у межах окремого набору даних, проте суттєво втрачають ефективність на реальних даних та щодо невідомих генеративних моделей. Виявлено наукову прогалину – відсутність комплексних мультимодальних засобів виявлення, здатних узгоджено аналізувати різнорідні сигнали, та обґрунтовано перспективність їх розроблення.

Ключові слова: виявлення фейкового контенту; штучний інтелект; дипфейк; трансформери; мультимодальний аналіз; глибоке навчання; інформаційна безпека.

Abstract

The theses present a review of modern methods and means of detecting fake content using artificial intelligence. A classification of approaches by four modalities – text, image, video, and audio – is proposed, with a cross-cutting multimodal analysis layer highlighted. Transformer-based detectors of fake and AI-generated texts, deepfake detection methods based on convolutional and transformer neural networks, and synthetic-speech detection systems using self-supervised learning are considered. It is shown that modern detectors achieve high accuracy within a single dataset but significantly lose effectiveness on in-the-wild data and against unseen generative models. A scientific gap is identified – the absence of comprehensive multimodal detection tools capable of jointly analysing heterogeneous signals – and the prospects for their development are substantiated.

Keywords: fake content detection; artificial intelligence; deepfake; transformers; multimodal analysis; deep learning; information security.

Вступ

Стрімкий розвиток генеративного штучного інтелекту докорінно змінив процес створення та поширення інформації. Якщо раніше виготовлення переконливих підробок вимагало спеціальних навичок і ресурсів, то сьогодні загальнодоступні генеративні моделі дають змогу за лічені секунди створювати правдоподібні фейкові тексти, зображення, відео та синтезоване мовлення. Це загострює проблему перевірки достовірності контенту: більшість користувачів не має змоги та часу самостійно верифікувати кожне повідомлення, а тому стає вразливою до маніпуляцій. З огляду на це розроблення методів та засобів автоматизованого виявлення фейкового контенту засобами штучного інтелекту є актуальним завданням інформаційної безпеки [1].

Попри значну кількість досліджень, присвячених виявленню підробок в окремих модальностях, комплексні мультимодальні засоби, здатні узгоджено аналізувати текст, зображення, відео та аудіо, залишаються малодослідженими. Метою дослідження є систематизація сучасних підходів до виявлення фейкового контенту засобами штучного інтелекту, характеристика наборів даних і метрик оцінювання, а також виявлення наукової прогалини та обґрунтування перспективного напряму подальших досліджень.

Результати дослідження

Проведений аналіз дозволив класифікувати методи виявлення фейкового контенту за чотирма основними модальностями – текст, зображення, відео та аудіо – з виокремленням наскрізного рівня мультимодального аналізу, що об'єднує різнорідні сигнали.

Виявлення фейкових текстів. Для цієї задачі широко застосовують донавчені трансформерні моделі (BERT, RoBERTa, ALBERT). На рівні цілісних новинних статей гібридні архітектури досягають високої якості – графово-доповнений ансамбль трансформерів забезпечує точність та F_1-міру близько 96,5 % на корпусі FakeNewsNet, тоді як на коротких твердженнях (набір LIAR) ефективність помітно нижча через брак контексту. Окремий напрям становить виявлення згенерованого штучним інтелектом тексту: метод DetectGPT використовує кривизну функції ймовірності перефразованих фрагментів для виявлення машинно-згенерованого тексту без додаткового навчання [2]. Застосовуються також підходи цифрового водяного знаку та статистичні детектори. Водночас надійність таких методів обмежена, оскільки перефразування здатне суттєво знизити їхню точність. Допоміжними ознаками слугують стилометричні характеристики та перевірка семантичної й фактологічної узгодженості.

Ефективність детекторів оцінюють за точністю (*accuracy*), влучністю (*precision*), повнотою (*recall*) та їхнім гармонічним середнім – F_1-мірою, а також за площею під ROC-кривою (AUROC):

$$F_1 = 2 * \frac{precision * recall}{precision + recall}. \quad (1)$$

Для систем виявлення аудіофейків основними метриками є рівень рівної помилки (EER) та мінімальна тандемна функція вартості виявлення (min t-DCF).

Виявлення фейкових зображень та відео. Базовим інструментом є згорткові нейронні мережі: детектор на основі архітектури Xception сягає точності близько 96 % на підмножині DeepFakes еталонного набору FaceForensics++ [3]. Розвиненими напрямами є аналіз у частотній області, методи на основі біологічних сигналів, детектори на основі візуальних трансформерів (ViT), а також виявлення зображень, синтезованих дифузійними моделями, зокрема за допомогою універсальних детекторів на базі CLIP. Ключовою проблемою залишається узагальнювальна здатність: за крос-наборового тестування показник AUC класичних детекторів падає приблизно до 65 % на наборі Celeb-DF, а на сучасному наборі реальних даних Deepfake-Eval-2024 ефективність найкращих відкритих моделей знижується на 45-50 % порівняно з лабораторними умовами, наближаючись до випадкового вгадування.

Для виявлення синтезованого мовлення та голосових підробок застосовують згорткові мережі на спектрограмах, наскрізні моделі типу RawNet2 і графові спектрально-часові мережі (AASIST), а також самокеровані звукові кодувальники (wav2vec 2.0, WavLM). У межах набору ASVspoof 2019 найкращі системи досягають рівня рівної помилки менш як 1 % [4]. Проте в умовах реальних даних узагальнювальна здатність різко погіршується: за результатами досліджень, EER зростає на 200–1000 %, а найкращі моделі у реальних умовах демонструють помилку близько 33 % [5].

Емпіричною основою досліджень слугують відкриті набори даних: для текстів – LIAR (12 836 тверджень), FEVER (185 445 тверджень), FakeNewsNet; для зображень і відео – FaceForensics++ (1 000 справжніх та 4 000 підроблених відео), Celeb-DF, DFDC (понад 128 тис. відеофрагментів); для аудіо – ASVspoof (набір для антиспуфінгу, видання 2019, 2021 та п'яте видання 2024 року), WaveFake (понад 100 тис. згенерованих аудіозаписів) та In-the-Wild (близько 38 год мовлення); для мультимодальних задач – Fakeddit (понад 1 млн зразків) та NewsCLIPpings.

На наскрізному рівні розвиваються методи перевірки крос-модальної узгодженості, виявлення контенту «поза контекстом» та детектори на основі мультимодальних великих мовних моделей. Узагальнення результатів дозволило виокремити ключові відкриті проблеми галузі: обмежену узагальнювальну здатність щодо невідомих генеративних моделей; вразливість до змагальних збурень і повторного ущільненя; постійне «протистояння» між генераторами та детекторами; складність ефективного злиття різномодальних ознак; недостатню пояснюваність ухвалених рішень.

Висновок

Проведений огляд підтверджує, що засоби штучного інтелекту забезпечують високу точність виявлення фейкового контенту в межах окремих модальностей та контрольованих наборів даних.

Водночас виявлено, що ключовою невирішеною проблемою залишається узагальнювальна здатність детекторів за умов реальних даних, а комплексні мультимодальні засоби, здатні узгоджено аналізувати текст, зображення, відео та аудіо, залишаються малодослідженими. Це обґрунтовує перспективність розроблення мультимодальних архітектур виявлення фейкового контенту, зокрема на основі агентних систем із застосуванням великих мовних моделей та зовнішніх інструментів, що становитиме напрям подальших досліджень.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Lin L. Detecting Multimedia Generated by Large AI Models: A Survey / L. Lin, N. Gupta, Y. Zhang et al. – arXiv preprint arXiv:2402.00045. – 2024. – DOI: 10.48550/arXiv.2402.00045.
2. Mitchell E. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature / E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn // Proceedings of the 40th International Conference on Machine Learning (ICML). – PMLR, 2023. – Vol. 202. – P. 24950–24962. – DOI: 10.48550/arXiv.2301.11305.
3. Rössler A. FaceForensics++: Learning to Detect Manipulated Facial Images / A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2019. – P. 1–11. – DOI: 10.1109/ICCV.2019.00009.
4. Wang X. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection / X. Wang, J. Yamagishi, M. Todisco et al. // Proceedings of Interspeech. – 2019. – P. 1008–1012. – DOI: 10.21437/Interspeech.2019-2249.
5. Frank J. WaveFake: A Data Set to Facilitate Audio Deepfake Detection / J. Frank, L. Schönherr // Proceedings of the NeurIPS Datasets and Benchmarks Track. – 2021. – DOI: 10.48550/arXiv.2111.02813.

Куперштейн Леонід Михайлович — к. т. н., доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця, email: kupershtein@vntu.edu.ua

Клименко Володимир Олександрович — аспірант групи F5-25аз, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: vovaklim2000@gmail.com

Leonid Kupershtein — PhD (eng), associated professor of information protection department, Vinnytsia National Technical University, Vinnytsia, email: kupershtein@vntu.edu.ua

Volodymyr Klymenko — Faculty of information technologies and computer engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: vovaklim2000@gmail.com