

СИСТЕМА ВИЯВЛЕННЯ ФЕЙКОВОГО МУЛЬТИМЕДІЙНОГО КОНТЕНТУ

¹Вінницький національний технічний університет

²ГО «ГВАРА МЕДІА»

Анотація

У роботі розроблено мультимодальну систему детекції фейкового контенту, що поєднує аналіз візуальної та аудіальної інформації. Відеоаналіз здійснюється за допомогою нейронної мережі ResNet50 з використанням підходу перенавчання класифікаційної надбудови для розпізнавання автентичних та підроблених відео. Для аналізу аудіо застосовано модель Wav2Vec, яка дозволяє оцінювати ймовірність синтетичного мовлення на рівні часових сегментів. Підсумкове рішення щодо автентичності контенту формується на основі диз'юнктивного правила прийняття рішень, згідно з яким мультимедійний об'єкт класифікується як фейковий у разі виявлення ознак підробки хоча б в одній із модальностей.

Ключові слова: діпфейк, мультимодальна детекція, нейронна мережа ResNet50, Wav2Vec, відеоаналіз, аудіоаналіз.

Abstract

This paper presents a multimodal system for fake content detection that combines the analysis of visual and audio information. Video analysis is performed using a ResNet50 neural network with a fine-tuned classification head for distinguishing between authentic and manipulated videos. Audio analysis is based on the Wav2Vec model, which enables the estimation of synthetic speech probability at the level of fixed time segments. The final decision on content authenticity is made using a disjunctive decision rule, according to which a multimedia object is classified as fake if signs of manipulation are detected in at least one of the modalities.

Keywords: deepfake, multimodal detection, neural network, ResNet50, Wav2Vec2, video analysis, audio analysis.

Вступ

Активний розвиток генеративних моделей суттєво спростило створення високоякісних діпфейків, які можуть переконливо імітувати зовнішність та голос людини. Такі матеріали становлять ризик для інформаційної безпеки, оскільки застосовуються для дезінформації, компрометації публічних осіб, шахрайських сценаріїв із підміною голосу та маніпуляцій у соціальних мережах. Особливу складність становлять випадки, коли фейк якісний і не має «очевидних» артефактів, тому перевірка лише однієї модальності (тільки відео або тільки аудіо) може бути недостатньою. У зв'язку з цим актуальним є створення практичних інструментів, що здатні автоматично виявляти ознаки синтетичності у відеоряді та мовному сигналі й формувати узгоджений висновок при їх спільному аналізі.

Результати дослідження

Розроблювана система виявлення фейкового мультимедійного контенту ґрунтується на результатах попередніх досліджень, присвячених аналізу можливостей нейронних мереж для детекції мультимедійних фейків, а також використанню великих мовних моделей у задачах автоматизованого фактчекінгу [1, 2]. У межах дослідження реалізовано прототип мультимодальної системи детекції, що поєднує аналіз обличчя у відеопотоці та аналіз синтетичності мовлення в аудіодоріжці. Відеогілка рішення базується на ResNet50 як надійному згортковому екстракторі ознак, використовується попередньо навчена основа, а для адаптації під задачу real/fake виконується тонке налаштування вихідних класифікаційних шарів нейромережі [3-5]. Такий підхід дозволяє швидко пристосувати модель до доменних спотворень (компресія, зниження якості, різні джерела відео) без повного перенавчання всієї мережі, зберігаючи корисні універсальні візуальні ознаки. Аудіогілка побудована на базі Wav2Vec, де модель виступає кодувальником мовного сигналу з подальшою класифікацією синтетичності. Аналіз виконується у вікнах фіксованої тривалості, що дає змогу стабільно обробляти як короткі, так і довгі записи та отримувати зведені показники для всього файлу [6], [7].

Практична цінність мультимодального підходу полягає в тому, що відео й аудіо містять різні типи слідів підробки. Відео частіше проявляє артефакти в зоні обличчя, міміки та дрібних текстур, тоді як аудіо відображає специфічні патерни синтезу голосу, надмірну регулярність і нетипову «гладкість» мовлення. Для підвищення узгодженості обробки аудіосигнал уніфікується до 16 кГц, що відповідає типовим вимогам pretrained-моделей Wav2Vec і зменшує варіативність ознак при надходженні даних з різних платформ і кодеків [7]. Підсумкова інтеграція реалізована у вигляді простого та інтерпретованого правила. Матеріал класифікується як «fake», якщо хоча б одна модальність перевищує свій поріг рішення. Така логіка обрана як більш безпечна для задачі виявлення підробок у прикладному сценарії, оскільки знижує ймовірність пропуску фейку у випадках, коли одна з модальностей «чиста», а інша містить ознаки синтетичності.

Система реалізована у вигляді вебзастосунку на основі фреймворку Streamlit [8]. В системі забезпечується завантаження файлів, запуск відео- та аудіоаналізу в окремих режимах або спільно, а також відображення зрозумілих користувачу результатів у вигляді зведеної оцінки та вердикту (рис. 1).

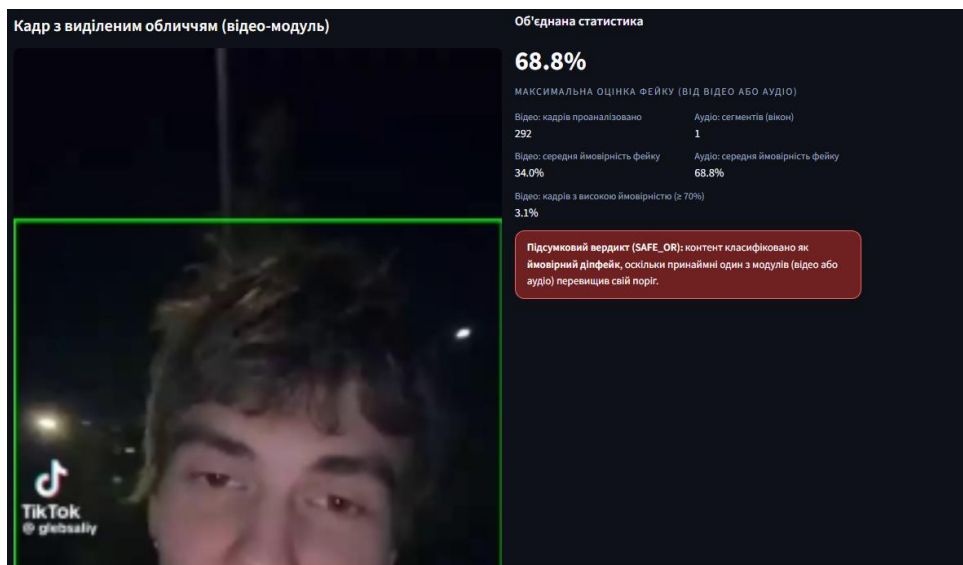


Рис. 1. Фрагмент відображення результатів мультимодального аналізу

Окремо передбачено налаштування порогів класифікації для відео та аудіо, що дозволяє адаптувати чутливість системи під умови експлуатації без зміни коду. Проведена експериментальна перевірка на прикладах різної складності показала, що система здатна реагувати як на «очевидні» дівфейки, так і на більш якісні генерації, а також коректно працювати з реальними матеріалами з відкритих джерел, забезпечуючи узгоджений висновок за рахунок спільного аналізу двох модальностей.

Висновки

Розроблена мультимодальна система виявлення фейкового мультимедійного контенту демонструє доцільність поєднання відео- та аудіоаналізу для підвищення ефективності детекції дівфейків у практичних умовах. Використання ResNet50 із фінтунінгом повноз'язного шару забезпечує адаптацію відеодетектора до доменних спотворень і різних джерел даних, тоді як аудіомодуль на базі Wav2Vec дозволяє фіксувати ознаки синтетичного мовлення через сегментний аналіз сигналу. Інтеграція результатів за правилом, при якому контент класифікується як фейковий, якщо хоча б одна з модальностей вказує на його підробленість, формує простий і прозорий механізм прийняття рішень, який знижує ризик пропуску підробки у випадках, коли хоча б одна з модальностей містить стійкі маркери синтетичності.

Водночас мультимодальний підхід має об'єктивні обмеження, пов'язані з якістю вхідних даних і умовами запису, сильна компресія, шум, перекриття обличчя або нестабільна аудіодоріжка можуть зменшувати впевненість окремих модулів і потребують коректного налаштування порогів чутливості. Загалом результати роботи підтверджують, що запропонована система є перспективною основою для прикладного застосування в задачах інформаційної безпеки та протидії дезінформації, поєднуючи практичну реалізацію у вебзастосунку зі зрозумілим механізмом інтеграції рішень і можливістю подальшого розвитку за рахунок розширення даних і вдосконалення моделей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Куперштейн Л.М., Сороколіт В.О., Прокопенко С.О. Аналіз можливостей великих мовних моделей для автоматизації фактчекінгу . Куперштейн Л. М., Сороколіт В. О., Прокопенко С. О. Аналіз можливостей великих мовних моделей для автоматизації фактчекінгу. // Матеріали Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2024)», 11-20 травня 2024 р. Електрон. текст. дані. 2024. URL: <https://conferences.vntu.edu.ua/index.php/mn/mn2024/paper/view/20855>.
2. Куперштейн Л.М., Людв Н. В., Прокопенко С. О. Аналіз можливостей нейронних мереж для виявлення мультимедійних фейків . Куперштейн Л. М., Людв Н. В., Прокопенко С. О. Аналіз можливостей нейронних мереж для виявлення мультимедійних фейків. // Матеріали Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2024)», 11-20 травня 2024 р. Електрон. текст. дані. 2024. URL: <https://conferences.vntu.edu.ua/index.php/mn/mn2024/paper/view/21749>.
3. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. arXiv, 2015. URL: <https://arxiv.org/abs/1512.03385> (дата звернення: 15.12.2025).
4. Transfer Learning for Computer Vision Tutorial. PyTorch Tutorials. URL: https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html (дата звернення: 15.12.2025).
5. timm (PyTorch Image Models): документація/моделі ResNet і pretrained-ваги. URL: <https://huggingface.co/docs/timm/index> (дата звернення: 15.12.2025)
6. Baevski A. et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv, 2020. URL: <https://arxiv.org/abs/2006.11477> (дата звернення: 15.12.2025).
7. Transformers: Wav2Vec2 (підготовка input_values через процесор/feature extractor, параметр sampling_rate). URL: https://huggingface.co/docs/transformers/model_doc/wav2vec2 (дата звернення: 15.12.2025).
8. Streamlit: st.file_uploader (завантаження файлів у вебзастосунок). URL: https://docs.streamlit.io/develop/api-reference/widgets/st.file_uploader (дата звернення: 15.12.2025).

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

Людв Назарій Вікторович — студент групи ІБС-24м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, email: nazarliudva@gmail.com

Прокопенко Сергій Олександрович – Керівник ГО «ІВАРА МЕДІА», email: prokopenko.serhii@gmail.com

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

Liudva Nazariy— student of group 1BS-24m, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: nazarliudva@gmail.com.

Prokopenko Serhii - Head of the NGO «GWARA MEDIA», email: prokopenko.serhii@gmail.com