УДК 004.056:004.8

Л. М. Куперштейн
А. В. Воєвода

ДОСЛІДЖЕННЯ МОЖЛИВОСТЕЙ ШІ ЧАТ-БОТІВ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВОГО ВІДЕОКОНТЕНТУ

Вінницький національний технічний університет

Анотація

У роботі проведено дослідження ефективності використання великих мовних моделей (LLM) чат-ботів загального призначення для виявлення фейкового відеоконтенту. Актуальність теми зумовлена стрімким поширенням дезінформації в соціальних мережах, що вимагає розробки доступних інструментів для швидкого фактчекінгу. У ході дослідження сформовано датасет, що включає 10 верифікованих фейкових відео та 10 автентичних матеріалів. Проаналізовано здатність різних ШІ-моделей (зокрема, Gemini та аналогів) ідентифікувати ознаки діпфейків, маніпулятивного аудіо та невідповідностей у візуальному ряді. Методологія дослідження передбачала сегментацію довгих відеофрагментів для оптимізації обробки в межах лімітів вхідних даних моделей. Результати роботи демонструють рівень точності та надійності чат-ботів у ролі інструментів первинної детекції підробок.

Ключові слова: діпфейк, великі мовні моделі, фактчекінг, детекція фейків, соціальні мережі, штучний інтелект.

Abstract

This study investigates the effectiveness of using general-purpose Large Language Model (LLM) chatbots for detecting fake video content. The relevance of this topic is driven by the rapid spread of misinformation on social media, which necessitates the development of accessible tools for rapid fact-checking. Throughout the research, a dataset was compiled consisting of 10 verified fake videos and 10 authentic materials. The study analyzes the ability of various AI models (including Gemini and its counterparts) to identify signs of deepfakes, manipulative audio, and visual inconsistencies. The research methodology involved segmenting long video clips to optimize processing within the models' input data limits. The results demonstrate the accuracy and reliability of chatbots as tools for initial forgery detection.

Keywords: deepfake, large language models, fact-checking, fake detection, social media, artificial intelligence.

Вступ

Стрімкий розвиток генеративного штучного інтелекту призвів до появи нових загроз у сфері інформаційної безпеки, зокрема, до масового поширення високоякісного фейкового відеоконтенту. Такі матеріали активно використовуються в соціальних мережах, таких як TikTok, YouTube, Telegram та ін. для проведення фішингових атак, поширення дезінформації та дестабілізації суспільної думки [1].

Основна проблема полягає в тому, що сучасні діпфейки стають дедалі переконливішими, імітуючи міміку, голос та поведінку реальних осіб. Це робить їх ефективним інструментом соціальної інженерії. Оскільки традиційні методи фактчекінгу потребують значного часу, доцільним є використання автоматизованих інструментів, зокрема інтелектуальних чат-ботів загального призначення на основі великих мовних моделей (LLM), для первинного виявлення та аналізу підозрілого контенту [2]. Метою дослідження є оцінювання ефективності чат-ботів на основі генеративного штучного інтелекту, як інструментів для оперативного виявлення фейків та маніпуляцій у відео.

Результати дослідження

Основою проведеного дослідження є практична перевірка того, як великі мовні моделі виявляють відеофейки та проводять автоматичний аналіз фактів [2, 3]. Для проведення тестування було

сформовано датасет, що складався з 20 мультимедійних об'єктів: 10 відео з підтвердженими ознаками маніпуляцій та 10 абсолютно автентичних матеріалів. Фейковий контент було зібрано шляхом моніторингу соціальних мереж, зокрема платформ TikTok та публічних каналів у Telegram, де більшість матеріалів мала фішинговий характер і містила маніпулятивні обіцянки соціальних виплат. Автентичні відео були відібрані з офіційних YouTube-каналів українських медіа та верифікованих сторінок державних установ. Технічний аналіз датасету показав, що файли переважно мали формати MP4 та MOV, вертикальну орієнтацію з роздільною здатністю від 464x848 до 1080x1920 пікселів, а їхній розмір варіювався від 5 до 18 мегабайт. Загальний хронометраж досліджуваних відеофайлів становив від 15 до 90 секунд.

Особливу увагу в методології було приділено подоланню технічних обмежень моделей щодо обсягу вхідних даних та тривалості завантажених файлів. Для цього застосовувався метод алгоритмічної сегментації: довгий файл чітко розбивався на рівноцінні фрагменти по 30 секунд. Кожен із цих сегментів завантажувався та аналізувався чат-ботом окремо за допомогою розробленого уніфікованого текстового запиту – промпту. Цей запит містив чіткі інструкції для проведення детального пошарового аналізу матеріалу на предмет часткових або повних фальсифікацій за чотирма ключовими критеріями: візуальний ряд, аудіодоріжка, текстовий вміст та цифрові сліди. Такий підхід дозволив виявити критично важливу закономірність: маніпуляція в соціальних мережах найчастіше має гібридний характер. Під час пофрагментного аналізу фіксувалися ситуації, коли перший сегмент відео містив абсолютно автентичний відеоряд і правдиву інформацію, виступаючи інструментом для здобуття довіри глядача. Проте вже у другому сегменті алгоритми фіксували непомітну підміну оригінальної аудіодоріжки на згенеровану ШІ, а в третьому виявляли накладення маніпулятивного тексту з фішинговим закликом до дії. Деталізований покроковий аналіз дозволив моделям точно локалізувати фейк і запобіг хибному схваленню всього відео через наявність у ньому реальних початкових кадрів.

Для оцінки ефективності розпізнавання ШІ-генерацій використовувалися просунуті та базові версії нейромереж від Google та OpenAI: моделі Gemini 3.1 Pro [4] та GPT-5.5 [5], а також базові конфігурації Gemini 3.1 Flash, Gemini 3.1 Flash-8B та оптимізована модель GPT-5.3 (використовується у безкоштовній версії ChatGPT).Важливим етапом дослідження стала перевірка здатності моделей до самостійного інтернет-фактчекінгу. Просунуті моделі (GPT-5.5 та Gemini 3.1 Pro) мали увімкнену функцію доступу до Інтернет, що дозволяло їм не лише проводити візуально-звуковий аналіз, але й автономно звертатися до мережі для звірки фактів (наприклад, перевіряти наявність реальних програм виплат від фондів). Базові конфігурації натомість працювали переважно в офлайн-режимі, спираючись виключно на свої попередньо навчені бази та аналіз метаданих.

Аналіз результатів фейкового контенту продемонстрував пряму залежність ефективності від класу моделі. Найвищу точність показали просунуті моделі GPT-5.5 та Gemini 3.1 Pro, які успішно ідентифікували 9 із 10 фейкових відео. Вони змогли синтезувати дані зі всіх шарів для виявлення прихованого фішингового наміру та успішно спростовували фейкові заяви завдяки пошуку в Інтернеті. Модель Gemini 3.1 Flash показала задовільний результат із точністю виявлення фейків на рівні 70% (7 із 10 відео), успішно знаходячи розбіжності між артикуляцією та звуковою доріжкою, проте маючи обмеження через відсутність глибокого веб-пошуку. Найнижчу ефективність продемонструвала базова безкоштовна модель GPT-5.3 (точність 40-50%). Під час тестування її підхід виявив критичну вразливість на етапі аудіоаналізу: модель повідомляла про недостатність даних для підтвердження підробки голосу та покладалася на поверхневі метадані файлу. Це свідчить про те, що базові алгоритми схильні до хибнонегативних результатів у випадках, коли візуальний ряд є автентичним оригінальним відеозаписом, а маніпуляція криється виключно в синтезованій аудіодоріжці. Схожі обмеження мала і базова модель Gemini 3.1 Flash-8B, яка швидко розпізнавала текстові накладення, але виявила низьку чутливість до інтонаційних аномалій.

Окремим вагомим аспектом дослідження стала перевірка реакції моделей на 10 автентичних відео з контрольної групи. Результати засвідчили, що флагманські моделі (GPT-5.5 та Gemini 3.1 Pro) досягли 100% точності, підтвердивши автентичність усіх 10 оригінальних відео. Базові версії також впоралися із цим завданням на високому рівні (точність GPT-5.3 на справжніх відео склала 90%). Алгоритми різних архітектур безпомилково ідентифікували природну міміку, узгоджене освітлення об'єктів у кадрі, а також наявність живих емоційних зітхань і пауз у голосі спікерів. Під час перевірки справжніх відео жодна з моделей не генерувала хибнопозитивних спрацьовувань, що доводить відсутність алгоритмічної упередженості.

Отримані результати підтверджують ключову тезу: процес верифікації відеоконтенту за допомогою штучного інтелекту є максимально доступним для пересічного користувача. Наявність заздалегідь підготовленого промпту нівелює необхідність у спеціальних технічних знаннях чи навичках програмування. Завдяки розробленій методиці сегментації та розумінню відмінностей між базовими і преміальними моделями, будь-яка людина може легко, швидко та ефективно перевірити підозріле відео на наявність гібридних маніпуляцій, просто завантаживши його у сучасний чат-бот невеликими частинами або невеликого розміру.

Висновки

Дослідження підтверджує, що великі мовні моделі є ефективним інструментом для швидкого виявлення відеофейків. Проте точність аналізу критично залежить від обраної моделі. Базові конфігурації (GPT-5.3, Gemini 3.1 Flash-8B) орієнтуються переважно на метадані та часто пропускають згенероване аудіо. Водночас просунуті версії (GPT-5.5, Gemini 3.1 Pro) здатні до глибокого семантичного мультимодального аналізу та успішно розпізнають фейки і маніпуляції у відеоповідомленнях.

Застосування методу алгоритмічної сегментації (розбиття відео на 30-секундні фрагменти) разом з уніфікованим промптом дозволяє обійти технічні обмеження чат-ботів. Це допомагає точно локалізувати гібридні підробки, де реальні кадри змішані з ШІ-вставками. Водночас експеримент виявив причину хибнонегативних результатів: базові конфігурації моделей часто пропускали маніпуляції через слабкість аналізу звукового шару. У випадках, коли відеоряд залишався повністю оригінальним, а підробка крилася виключно в якісно згенерованому ШІ голосі, такі алгоритми покладалися на поверхневі метадані контейнера файлу і помилково вважали матеріал справжнім. При цьому апробація на суто автентичних відеоматеріалах показала, що за правильного налаштування промпту алгоритми працюють об'єктивно і не видають хибнопозитивних спрацьовувань.

У системі інформаційної безпеки LLM-асистенти виконують функцію першої лінії захисту, тоді як попіксельну експертизу варто залишати вузькоспеціалізованим програмам. Основна перевага запропонованого підходу — його доступність. Будь-який користувач без навичок програмування може швидко перевірити підозріле відео за допомогою смартфона. Це робить таку методику дієвим та масовим інструментом для підвищення цифрової гігієни та протидії дезінформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Vaccari C. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news / C. Vaccari, A. Chadwick // Social Media + Society. – 2020. – Vol. 6, № 1. – P. 1–13.
2. Куперштейн Л. М. Система виявлення фейкового мультимедійного контенту / Л. М. Куперштейн, Н. В. Людва, С. О. Прокопенко // Наукові праці Вінницького національного технічного університету. – Вінниця : ВНТУ, 2024.
3. Аналіз можливостей великих мовних моделей для автоматизації фактчекінгу [Електронний ресурс] / Л. М. Куперштейн, В. О. Сороколіт, С. О. Прокопенко // Матеріали Всеукраїнської науково-практичної інтернет-конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2024)». – Вінниця : ВНТУ, 2024. – Режим доступу: https://conferences.vntu.edu.ua/index.php/mn/mn2024/paper/view/20855.
4. Gemini: A Family of Highly Capable Multimodal Models [Електронний ресурс] / Gemini Team, Google DeepMind // arXiv. – 2023. – Режим доступу до ресурсу: https://arxiv.org/abs/2312.11805.
5. GPT-4 Technical Report [Електронний ресурс] / OpenAI // arXiv. – 2023. – Режим доступу до ресурсу: https://arxiv.org/abs/2303.08774.

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

Воєвода Аліна Віталіївна — студентка групи 1БКС-24б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, email: alinavoievoda77@gmail.com

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

Voievoda Alina — student of group 1BKS-24b, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: alinavoievoda77@gmail.com.