УДК 004.93

В.М. Білоус[1]
О.В.Кириченко[2]

ВИЯВЛЕННЯ ВІЗУАЛЬНОЇ ГЕНЕРОВАНОЇ ДЕЗІНФОРМАЦІЇ У ВИГЛЯДІ МЕМІВ ЗОБРАЖЕНЬ В СОЦІАЛЬНИХ МЕРЕЖАХ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ

[1]Вінницький Національний технічний університет
[2]Вінницький інститут конструювання одягу і підприємництва

Анотація

У роботі розглядається проблема поширення візуальної дезінформації у вигляді мемів-зображень у соціальних мережах, згенерованих за допомогою штучного інтелекту. Особливу увагу приділено застосуванню методів комп'ютерного зору та алгоритмів глибокого навчання для виявлення аномальних ознак у зображеннях, а також аналізу текстових компонентів мемів. Запропоновано концепцію комплексного аналізу мультимодального контенту, що включає зображення, текст і контекст поширення. Наведені результати є основою для побудови автоматизованих систем моніторингу та фільтрації фейкового контенту, що сприятиме підвищенню рівня інформаційної безпеки користувачів.

Ключові слова: штучний інтелект, зображення-меми, соціальні мережі, комп'ютерний зір, генеративні моделі, глибоке навчання.

Вступ

Останніми роками соціальні мережі стали основним каналом комунікації та обміну інформацією, зокрема у візуальному форматі. Меми-зображення – короткий візуально-текстовий контент із сильним емоційним або ідеологічним навантаженням, що часто використовуються для маніпуляції громадською думкою. З появою потужних генеративних моделей, таких як GAN, DALL-E, Stable Diffusion тощо, можливості створення фейкового візуального контенту значно розширилися. Проблема полягає в тому, що традиційні методи перевірки фактів не здатні ефективно обробляти мультимодальний контент. Це обумовлює необхідність розробки автоматизованих рішень на базі штучного інтелекту, орієнтованих на виявлення згенерованих мемів та їх семантичний аналіз.

Основна частина

Для розв'язання поставленої задачі пропонується трирівневий підхід:

- аналіз зображень, який використовує застосування моделей комп'ютерного зору (CLIP, ResNet, EfficientNet) для детектування артефактів генерації, такі як неприродні текстури, спотворення облич, неузгодженість світлотіней тощо;
- текстовий аналіз, який передбачає використання трансформерних моделей (BERT, RoBERTa) для виявлення ознак емоційної маніпуляції, пропаганди та семантичних розбіжностей між зображенням і підписом;
- контекстуальний аналіз, при якому проводиться моделювання поведінки поширення контенту (метаінформація, бот-активність, часові шаблони) з метою виявлення організованих інформаційних кампаній.

Покроковий опис процесу роботи запропонованої системи (рис.1):

1. Отримання мемів, які опубліковані в соціальних мережах з можливим вбудованим текстом.
2. Аналіз зображення моделями комп'ютерного зору для виявлення аномалій, характерних для генеративного ШІ (некоректна анатомія, неприродне освітлення, дефекти текстур), аналіз метаданих зображення (якщо доступні EXIF-дані), порівняння з базами зображень для виявлення підробок або редагувань.
3. Аналіз вбудованого тексту технологіями OCR (оптичне розпізнавання тексту) з використанням трансформерних NLP-моделей для розпізнання тексту із зображення з

семантичним аналізом тексту на наявність емоційних, маніпулятивних або неправдивих тверджень і суперечностей між зображенням і текстовим повідомленням.

4. Контекстуальний аналіз поведінки у соціальних мережах (API моніторингу, бот-трекінг, часові шаблони) для визначення джерела публікації (чи є джерело авторитетним, чи пов'язане з підозрілими акаунтами), аналіз патернів поширення (бот-активність, масове ретвітення, синхронізація з іншими акаунтами), перевірка на залученість до відомих дезінформаційних кампаній.

5. Інтеграція результатів у модель ШІ, яка приймає на вхід дані з попередніх етапів і на основі всіх ознак (візуальних, текстових, контекстуальних) може класифікувати мем як щирий (гумористичний або інформативний контент без ознак дезінформації), або як згенеровану дезінформацію (фейковий, маніпулятивний або пропагандистський контент).

Рисунок 1 – Процес роботи системи виявлення візуальної генерованої дезінформації у вигляді мемів

Об’єднання візуальних і текстових ознак підвищує точність класифікації дезінформаційного контенту у порівнянні з одноканальними підходами. Більшість згенерованих мемів мають типові шаблони структури, що дозволяє будувати евристичні правила попередньої фільтрації.

Висновок

Інтеграція методів штучного інтелекту у процес виявлення дезінформації у вигляді зображень-мемів є перспективним напрямком для побудови сучасних систем цифрової безпеки. Запропонований підхід є ефективним у виявленні згенерованого контенту на основі аналізу зображення, тексту та мережевого контексту. Подальші дослідження повинні бути спрямовані на вдосконалення мультимодальних моделей, а також на розробку візуальних індикаторів достовірності контенту для кінцевих користувачів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Goodfellow I. et al. Generative Adversarial Nets // Advances in Neural Information Processing Systems. – 2014.
2. Radford A. et al. Learning Transferable Visual Models From Natural Language Supervision // ICML. – 2021.
3. Zhang H. et al. Detecting GAN-generated Imagery using ResNet Features // arXiv preprint. – 2019.
4. Zellers R. et al. Defending Against Neural Fake News // NeurIPS. – 2019.

Білоус Віталій Михайлович, асистент кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, місто Вінниця, vitalii.bilous@vntu.edu.ua

Кириченко Олександр Вікторович, директор Фахового коледжу, Вінницький інститут конструювання одягу і підприємництва, місто Вінниця, olkirichenko@gmail.com

DETECTION OF VISUALLY GENERATED DISINFORMATION IN THE FORM OF MEMES IMAGES IN SOCIAL NETWORKS BASED ON ARTIFICIAL INTELLIGENCE

Abstract

The article examines the problem of the spread of visual disinformation in the form of meme images in social networks generated using artificial intelligence. Particular attention is paid to the use of computer vision methods and deep learning algorithms to detect anomalous features in images, as well as the analysis of text components of memes. The concept of a comprehensive analysis of multimodal content, including images, text, and the context

of distribution, is proposed. The results presented are the basis for building automated systems for monitoring and filtering fake content, which will contribute to increasing the level of information security of users

Keywords: artificial intelligence, image memes, social networks, computer vision, generative models, deep learning.

Bilous Vitalii Mykhailovych, Assistant of the Department of Management and Information Systems Security, Vinnytsia National Technical University, Vinnytsia, vitalii.bilous@vntu.edu.ua

Kyrychenko Oleksandr Viktorovych, Director of the Vocational College, Vinnytsia Institute of Clothing Design Entrepreneurship, Vinnytsia, olkirichenko@gmail.com