УДК 004.89

ЗАСТОСУВАННЯ КЛАСИФІКАЦІЇ ШІ-ГЕНЕРОВАНИХ ПОВІДОМЛЕНЬ В ІНТЕРНЕТ-БРАУЗЕРАХ ДЛЯ БОРОТЬБИ З ПРОПАГАНДОЮ

Іванова О.С.
Науковий керівник — д.т.н., проф. Терзіян В.Я.
Харківський національний університет радіоелектроніки, каф. ШІ,
м. Харків, Україна
e-mail: oleksandra.ivanova@nure.ua

This work explores the intricate dynamics of propaganda and AI convergence in the contemporary digital era. It highlights the challenges arising from the rapid dissemination of information through digital platforms, blurring the lines between trustworthy news and propaganda. It proposes using a combination of GAN and classification models to detect and combat both human and AI-generated propaganda, advocating for their integration into web browsers for real-time identification of propagandistic content. This approach aims to enhance media literacy and critical thinking skills among users, thereby mitigating the harmful effects of propaganda on public discourse and decision-making processes, ultimately strengthening democratic societies.

У сучасну цифрову епоху конвергенція штучного інтелекту (ШІ) і пропаганди започаткувала нову еру, що характеризується складною соціально-технічною динамікою. Ера швидкого поширення інформації через цифрові платформи робить виклик розрізнення між надійними новинами (постами, історіями, тощо) та пропагандою все більш складним. Беручи за основу фільтрацію Інтернет-зображень для розмиття потенційно травмуючого контенту з можливістю коригування налаштувань їх відображення у браузері, я пропоную у такий же самий спосіб застосувати поєднання генеративних змагальних мереж (англ. GANs) із класифікацією трансформером для виявлення як людської, так і створеної ШІ пропаганди.

Пропаганда, що визначається як поширення упередженої або оманливої інформації з наміром маніпулювати громадською думкою, діє за допомогою переконливих повідомлень, вибіркової презентації та маніпулювання емоціями. Її виявлення є важким завданням навіть для людини, особливо при критично низькому рівні медіаосвіти сучасної світової громадськості. З появою цифрових медіа поширення пропаганди досягло небачених рівнів. Соціальні мережі, зокрема, стали місцями для швидкого поширення дезінформації та пропаганди, сприяне алгоритмами, спрямованими на максимізацію залученості користувачів. Як результат, це призвело до зниження рівня довіри інформації з офіційних джерел, при цьому підвищуючи впевненість у розповсюдженій не експертній думці стосовно конкретних питань по всьому світу. Це стало можливим і завдяки створенню і публікації багатьох теорій у соціальних мережах, з доказами і

без, що дають можливість вважати довіру будь-якій офіційній інформації нелогічною. Людям сучасності набагато легше повірити невідомому їм користувачу в Інтернеті, який грає роль «такого ж самого обуреного чи стурбованого громадянина», не підозрюючи, що ця маніпуляція виступає проти них самих. Так само в сучасності відомі випадки, коли відома сторона (людина, компанія, тощо), що прямо не бере участь у розглядаємих питаннях, неодноразово публічно викладає свою думку, що комплементує інформації, раніше визначеній як пропаганда. Користуючись своїм статусом, ця сторона нав'язує корисну пропаганді думку людям, що є її послідовниками. Ефекти довіри відомій чи, навпаки, випадковій людині чи теорії через велику популярність чи число прочитань замість офіційних джерел і доступних доказів допомагають ширити пропаганду з ще більшою ефективністю ніж коли-небудь раніше.

З появою відкритих ШІ сервісів, таких як ChatGPT, пропаганда отримала собі потужний прилад для генерування ШІ контенту, який в багатьох випадках може навіть перевищити людський за ознаками приємності та переконливості. За останні роки були виявлені випадки використання генеративних мереж для створення фейкових зображень та відео (англ. Deepfake) задля просування конкретних ідей або емоцій, залучення громадськості діяти певним чином, який був би корисний стороні, що застосовує пропаганду. Однак якщо поки що є ознаки, по яким середньостатистична людина може сама відрізнити візуальну синтетичну пропаганду від реальної події чи людини (спадання маски при поворотах, якість зображення, тощо), то з синтетичними текстами це зробити складніше. Коли зображення чи відео лише існують, у текста є час переконати свого читача у своєму контенті. Наприклад, було визначено, що ChatGPT може писати дуже красиві та переконливі тексти, у яких пропаганда накладена тонким шаром, роблячи роботу класифікації як пропаганди складніше для людини.

На тлі цього ШІ виступає також як потужний інструмент у боротьбі з пропагандою. Алгоритми машинного навчання здатні аналізувати величезні обсяги даних та помічати складні представлення ознак пропаганди, пропонуючи потенціал автоматизування виявлення пропагандистського контенту. Таким чином, ми увійшли у еру, де ШІ змагається з ШІ.

Одним з перспективних підходів до використання ШІ в боротьбі з пропагандою є інтеграція алгоритмів виявлення прямо в Інтернет-браузери. Так користувачі можуть отримувати в реальному часі попередження або індикатори під час доступу до потенційно пропагандистського контенту. Цей активний підхід дасть людям змогу приймати обмірковані рішення щодо інформації, яку вони споживають, сприяючи медіаосвіті та навичкам критичного мислення. Проте однією з основних перепон цієї мети є необхідність міцних, безпечних і точних моделей виявлення, які зможуть

надійно відрізняти пропаганду від звичайних новин чи постів, навіть якщо вона була створена ШІ.

Для цього пропонується створити систему поєднання генеративних та класифікуючих моделей. Вважаючи, що існує датасет конкретних пропагандистських текстів, його буде застосовано для навчання генеративної моделі, де генератор буде створювати повідомлення, намагаючись імітувати наведені у датасеті, а критик — оцінювати їх, поки не зможе помітити різниці у сенсі тексту. Таким чином, буде отриманий датасет згенерованої ШІ пропаганди, такого ж розміру, що й початковий датасет людської пропаганди. Об’єднавши ці два датасета в один та перемішавши записи, треба лише створити класифікатор. Для цього можна застосувати вже навчену модель (бажано трансформер через високі показники точності при роботі з текстовими даними, наприклад, GPT) та налаштувати її за допомогою змішаного датасету пропаганди для бінарної класифікації на класи «Пропаганда» чи «Чистий текст». Таким чином, завдяки відкриттю до людських можливостей написання пропаганди та ШІ генерованих подібних повідомлень, фінальна модель має бути набагато більш чутлива до будь-якого тексту, що містить у собі пропаганду, навіть приємно та гарно обставленого, як може писати модель GPT.

Наступним кроком пропонується застосувати подібну модель не для закритих зборів аналітики, але для попередження людей у медіапросторі в реальному часі. Інтеграція подібної моделі ШІ для виявлення пропаганди в Інтернет-браузерах буде мати глибокі позитивні наслідки для медіаосвіти та демократії. Забезпечуючи користувачів інструментами для відрізняння надійної інформації від пропаганди, суспільство може пом'якшити шкідливі наслідки пропаганди на громадський діалог та процеси прийняття рішень. Як і при фільтрації сумнівних зображень, користувач буде мати змогу прочитати пост чи новину, що не буде порушувати його прав, але повідомлення від браузера про виявлення пропаганди у поданому тексті все ж допоможе користувачам розвивати культуру критичного мислення та сумніву до джерел медіа, що, в свою чергу, зміцнить основи демократичних суспільств.

Отже, інтеграція алгоритмів ШІ в Інтернет-браузерах для виявлення пропаганди є перспективним кроком у вирішенні питання поширення пропаганди в цифрову епоху, яка не вилучить сумнівний контент з уваги людей, але навчить їх розрізняти його самим, зміцнюючи суспільство.

Список використаних джерел

1. Propaganda / editors: Jackall R. 1995. 449

2. Goldstein J. A., Chao J., Grossman Sh, Stamos A., Tomz M. How persuasive is AI-generated propaganda? *PNAS Nexus*. 2024. Vol. 3, No 2. P. 1-7

3. Jones D. G. Detecting Propaganda in News Articles Using Large Language Models. *Engineering: Open Access*. 2024. Vol. 2, No 1. P. 1-12.