Технології розпізнавання дезінформації в соціальних мережах

Анна Шелестова
Кафедра цифрових комунікацій та інформаційних технологій
Харківська державна академія культури
Харків, Україна
anna_shelestova@ukr.net

Abstract. This paper explores the evolution of technologies for detecting disinformation on social media platforms. As digital misinformation threatens public discourse and democratic processes, advanced technological solutions have become essential. The field has progressed from traditional media monitoring to sophisticated systems employing artificial intelligence, machine learning, neural networks, and automated fact-checking tools that can analyze vast datasets to identify misleading content in real-time. Despite these advancements, significant challenges persist regarding ethical implications, potential freedom of expression infringements, and algorithmic bias. The paper emphasizes that effective disinformation management requires collaboration between technology developers, policymakers, and academic institutions to create balanced approaches that combat misinformation while preserving civil liberties.

Ключові слова: дезінформація, соціальні медіа, штучний інтелект, перевірка фактів, регулювання.

Наприкінці 20-го століття виникли дискусії щодо впливу влади та управління суспільством, особливо під впливом робіт таких філософів, як Мішель Фуко. Його теорії стверджували, що динаміка влади в суспільстві формує спосіб управління та контролю інформації, закладаючи фундаментальні ідеї для розуміння управління інформацією в цифровому просторі (1). Коли онлайн-платформи почали домінувати, проблема дезінформації ставала все більш очевидною, підкреслюючи потребу в нових теоретичних засадах та інструментах для боротьби з цим явищем (2, 3).

Розвиток цифрових технологій у 2000-х роках створив як можливості, так і виклики для розпізнавання дезінформації. Платформи соціальних мереж уможливили швидке поширення неправдивої інформації, що призвело до широкомасштабних соціальних наслідків, у тому числі до підриву довіри до усталених інституцій (4). У цей період з'явилися автоматизовані системи, призначені для відстеження та виявлення патернів дезінформації, причому перші моделі зосереджувалися на поведінці розповсюджувачів інформації, а не на самому контенті (5, 6). З поглибленням розуміння механізмів дії дезінформації, особливо в контексті важливих політичних питань, науковці та практики визнали, що ефективні методи виявлення дезінформації потребують багатогранного підходу (технологічні рішення, контекстуальні нюанси дезінформаційних кампаній) (3, 7). Розвиток технологій штучного інтелекту (ШІ) та машинного навчання (ML) ще більше трансформував цифровий ландшафт, що призвело до появи складних інструментів, здатних розпізнавати діпфейки та інші маніпулятивні медіа-форми (8, 9).

На початку 2020-х років нагальна потреба боротися з дезінформацією в соціальних мережах спонукала до співпраці між різними секторами, зокрема державними установами, технологічними компаніями та академічними установами. Ці партнерства були спрямовані на створення комплексних механізмів для розпізнавання та зменшення впливу дезінформації, а також на навчання громадськості щодо важливості критичного ставлення до онлайн-контенту (4, 10). Оскільки технологічний ландшафт продовжує розвиватися, зростає і потреба в інноваційних рішеннях для боротьби з дезінформацією та підвищення стійкості демократичних суспільств до дезінформації.

ML стало життєво важливим компонентом у боротьбі з дезінформацією. Аналізуючи величезні масиви даних, алгоритми ML можуть виявляти закономірності, які відрізняють достовірний контент від неправдивого. Такі методи, як обробка природної мови (NLP), використовуються для виявлення лінгвістичних нюансів, пов'язаних з дезінформаційними текстами, тим самим підвищуючи точність систем виявлення фейкових новин (11, 12). Такі алгоритми, як дерева рішень і нейронні мережі, виявилися особливо ефективними в розпізнаванні оманливих наративів, враховуючи контекстуальні зв'язки між авторами, темами і ключовими словами (13, 12).

Графові нейронні мережі (GNN) представляють собою якісний підхід до виявлення дезінформації шляхом аналізу взаємозв'язку даних у соціальних мережах. Цей метод фіксує складні взаємозв'язки між

користувачами, контентом і потоками інформації, що дозволяє вживати проактивних заходів для припинення поширення такого контенту (13).

Автоматизовані інструменти перевірки фактів оцінюють твердження, що містяться в публікаціях у соціальних мережах, за перевіреними базами даних і надають негайний відгук про їхню достовірність, що робить їх цінним активом у боротьбі з дезінформацією на великих соціальних платформах, таких як YouTube та Instagram (12, 14).

Окрім автоматизованих технологій, все більшої популярності набувають громадські ініціативи. Платформи соціальних мереж дедалі більше покладаються на краудсорсингову перевірку фактів, де користувачі можуть маркувати контент на основі його точності (14).

Отже, співпраця між розробниками технологій, політиками та науковими установами матиме вирішальне значення для створення ефективних стратегій, які не лише боротимуться з дезінформацією, але й навчатимуть користувачів критично ставитися до контенту. Майбутні перспективи цих технологій залежать від їхнього етичного застосування та адаптивності нормативно-правової бази до викликів, що постають перед цифровими ландшафтами, які невпинно змінюються.

Література

[1] Kyza, E.A., Varda, C., Panos, D., Karageorgiou, M., Komendantova, N., Perfumi, S.C., Shah, S.I.H., & Hosseini, A.S. (2020). Combating misinformation online: re-imagining social media for policy-making. Internet Policy Review, 9(4). https://doi.org/10.14763/2020.4.1514

[2] Christine Clark. (2024, September 17). How AI can help stop the spread of misinformation. UC San Diego. https://today.ucsd.edu/story/how-ai-can-help-stop-the-spread-of-misinformation

[3] Noah Pflueger-Peters. (2022, May 20) An algorithm to detect fake news. UC Davis College of Engineering. https://engineering.ucdavis.edu/news/algorithm-detect-fake-news

[4] Chandrasekaran R., Sadiq T. M., Moustakas E. (2024) Racial and demographic disparities in dusceptibility to health misinformation on social media: national survey-based analysis. J Med Internet Res, 26. doi: 10.2196/55086

[5] Jia, C. & Lee, T. (2024). Journalistic interventions matter: understanding how Americans perceive fact-checking labels. Harvard Kennedy School (HKS) Misinformation Review, 5(2). https://doi.org/10.37016/mr-2020-138

[6] David Adam. (2025, January 15). Does fact-checking work? Here's what the science says. Scientific American. https://www.scientificamerican.com/article/does-fact-checking-work-on-social-media/

[7] States United Domocracy Center. (2025). Social Media Policies: mis/disinformation, threats, and harassment (2025). https://statesunited.org/resources/social-media-policies/

[8] Mauro Fragale & Valentina Grilli (2024). Deepfake, deep trouble: the European AI act and the fight against AI-Generated Misinformation. Columbia Journal of European Law. https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/?cn-reloaded=1

[9] Pilati F., Venturini T. (2025) The use of artificial intelligence in counter-disinformation: a world wide (web) mapping. Front. Polit. Sci. 7:1517726. doi: 10.3389/fpos.2025.1517726

[10] Kristina Hook & Ernesto Verdeja (2022). Social media misinformation and the prevention of political instability and mass atrocities. Human Rights & IHL. https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/

[11] Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, & Ido Wulkan. (2021). AI and the Future of Disinformation Campaigns. Center for Security and Emerging Technology. https://doi.org/10.51593/2021CA011

[12] American's Cyber Defense Agency. (2022). Tactics of Disinformation https://www.cisa.gov/sites/default/files/publications/tactics-of-disinformation_508.pdf

[13] Lan, D. H., & Tung, T. M. (2024). Exploring fake news awareness and trust in the age of social media among university student TikTok users. Cogent Social Sciences, 10(1). https://doi.org/10.1080/23311886.2024.2302216

[14] Amanda Hetler. (2025). 11 ways to spot disinformation on social media. TechTarget. https://www.techtarget.com/whatis/feature/10-ways-to-spot-disinformation-on-social-media