UDC 004.738.5:316.4

Oleksandr Dorenskyi, Viktor Kolesnyk
dorenskyiop@kntu.kr.ua, kolesnykviktor2006@gmail.com
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine

AI-BASED FACT-CHECKING METHOD FOR COUNTERING DISINFORMATION IN CYBERSPACE

In today's information space, the issue of disinformation is extremely significant. Due to information overload, it is often difficult for individuals to identify disinformation on their own. This issue has become especially critical amid Russian aggression. Under these circumstances, artificial intelligence technologies serve as powerful tools both for generating and detecting disinformation within the digital information environment.

According to the International Economic Forum (2024), disinformation has ranked first among short-term global risks. Generative AI is responsible for up to 30% of fake content on social media platforms, while 50% of users cannot distinguish deepfakes from authentic materials. Key trends include increased automation, personalization of attacks, and hybrid campaigns [1-3].

The research presented in [4] and [5] addresses the pressing issue of improving the effectiveness of counter-disinformation efforts within information warfare, which in Ukraine is carried out by the Center for Countering Disinformation. However, in the coordination model of the information warfare system presented in [4], little attention is given to fact-checking tasks and the use of AI technologies to counter disinformation.

AI technologies, particularly NLP (Natural Language Processing) and automatic fact-checking, demonstrate a high level of effectiveness in detecting fake information. Models such as BERT and GPT-4 analyze semantic patterns and sentence structures, achieving accuracy levels of up to 95% when integrated with SVM systems [6]. Computer vision is also actively utilized in combating deepfakes. Tools such as Sensity AI employ video forensics to detect artifacts like inconsistent blinking or anatomical errors. These technologies help to analyze the spread of fake content on social networks by examining geographical and temporal patterns [7].

It is also important to highlight hybrid human-machine systems, which combine AI algorithms with expert verification. The UK's Full Fact system integrates NLP models with human fact-checking of political claims [2]. In Ukraine, a similar approach is employed by the Trusted Flaggers organization.

In the research [5], a model of an information warfare system for the National Security and Defense Coordination Center has been developed. According to the authors, this system should include a monitoring platform integrating text classification algorithms to analyze disinformation trends. Creating an open annotated database of fake content, similar to EUvsDisinfo, is considered crucial. Decentralized systems, particularly blockchain-based solutions, could form the foundation for independent fact-checking databases. Predictive algorithms, such as MIT's Veracity, are also effective, anticipating the virality of fake content several hours before it spreads widely [3].

For Ukraine, which has been subjected to an ongoing information war for many years, the implementation of AI tools for detecting disinformation is, according to the authors, a matter of national security. To effectively counter disinformation, it is necessary to adopt a comprehensive approach that actively integrates various artificial intelligence technologies, tools, and platforms.

References

1. RAND: Towards an AI-Based Counter-Disinformation Framework: Website. URL: www.rand.org/pubs/commentary/2021/03/towards-an-ai-based-counter-disinformation-framework.html (Accessed 04.04.2025).

2. Frontiers: The use of artificial intelligence in counter-disinformation: Website. URL: www.frontiersin.org/journals/political-science/articles/10.3389/fpos.2025.1517726/full (Accessed 04.04.2025).

3. Sedova K., McNeill Ch., Johnson A., Joshi A., Wulkan I. AI and the Future of Disinformation Campaigns : CSET Policy Brief. 2021. URL: https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Disinformation-Campaigns-Part-2.pdf (Accessed 04.04.2025).

4. Доренський О.П. Модель поведінки держави в умовах проявів ознак інформаційної експансії, агресії, війни. *Інформаційна безпека держави, суспільства та особистості*. 2015. С. 131-133.

5. Доренський О.П., Улічев О.С., Задорожний К.О., Коваленко А.С., Дрєєва Г.М. Концептуальна модель системи інформаційного протиборства координаційного центру з питань національної безпеки і оборони. *Центральноукраїнський науковий вісник. Технічні науки*. 2024. Вип. 10(41). Ч. 2. С. 23-31.

6. Ashiq Mohammed M., Pradeesh E., Jeevanantham M., Anandhu B.S., Fake News Classification Using Machine Learning. *International Journal of Engineering Research & Technology (IJERT)*. 2023. Volume 11, Issue 03. DOI : 10.17577/IJERTCONV11IS03020 (Accessed 03.04.2025).

7. The Trusted Web: 13 AI-Powered Tools for Fighting Fake News: Website. URL: https://thetrustedweb.org/ai-powered-tools-for-fighting-fake-news/ (Accessed 03.04.2025).