UDC 004.8:004.912:316.774 DOI: 10.31673/2412-9070.2026.318102

M. HNATYSHYN, PhD Student;
ORCID: 0009-0009-0813-3602
O. NEDASHKIVSKIY, DSc (Engin.), Prof.,
ORCID: 0000-0002-1788-4434
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

ADVANCED ARTIFICIAL INTELLIGENCE ARCHITECTURES FOR THE PROACTIVE SEARCH AND CLASSIFICATION OF MANIPULATIVE INFORMATION IN DIGITAL DISCOURSE

The subject of this research is the development of an automated framework for the detection and classification of manipulative information within digital ecosystems, leveraging hybrid machine learning architectures and explainable artificial intelligence (XAI). In an era characterized by the proliferation of computational propaganda and generative AI-driven disinformation, traditional reactive detection methods are increasingly insufficient. This article introduces a novel, multi-layered architecture — the Cognitive-Linguistic Manipulation Analysis Framework (CLMAF) — designed to identify non-transparent influence attempts by integrating linguistic pragmatics, stylometric profiling, and cross-modal consistency checking. The primary objective is to enhance the interpretability of detection models, thereby fostering user trust and enabling proactive moderation of harmful content. The methodology employs an ensemble of transformer-based models (specifically fine-tuned BERT and RoBERTa architectures) integrated with Graph Neural Networks (GNN) to analyze both the semantic content and the structural propagation patterns of potential manipulation. The scientific novelty lies in the synthesis of a stakeholder-centric, multimodal architecture that moves beyond binary veracity classification toward a nuanced identification of psychological manipulation techniques. Findings suggest that the integration of XAI not only improves the transparency of AI-driven decisions but also enhances the overall robustness of the system against adversarial attacks. The proposed framework effectively bridges the gap between high-performance "black-box" neural networks and the necessity for human-centric accountability in information security. Future research directions include the empirical validation of the CLMAF architecture against evolving generative threats and the refinement of cross-lingual manipulation markers. The text further elaborates on the mathematical foundations of the verification probability model and the similarity-based retrieval mechanism, providing a comprehensive blueprint for next-generation information defense systems.

Keywords: manipulative information, artificial intelligence, machine learning, explainable AI (XAI), computational linguistics, digital disinformation, transformer models, neural network architectures, software engineering, data processing.

Introduction

The global information landscape is currently undergoing a profound transformation, driven by the rapid evolution of digital communication technologies and the pervasive integration of artificial intelligence into the production and dissemination of content [1]. While these advancements have democratized access to information, they have also facilitated the rise of sophisticated manipulative operations that threaten the integrity of public discourse and democratic processes [2]. Manipulation, distinguished from transparent persuasion by its non-transparent nature, has become a primary tool for both state and non-state actors [3]. The complexity of detecting manipulative information is exacerbated by the "black-box" nature of contemporary machine learning models, where internal reasoning remains opaque to human observers.

Previous research by the authors has emphasized the necessity of using neural network-based methods to enhance the reliability of information on the internet [7, 15]. Specifically, Hnatyshyn and Nedashkivskiy have proposed conceptual frameworks for integrating XAI within fake news detection systems to enhance transparency and user agency [5]. Building upon this foundation, the current study seeks to develop a more granular architectural solution specifically tailored for the automatic search and identification of manipulative information. Unlike broad misinformation detection, which focuses primarily on factual veracity, manipulation detection requires an understanding of rhetorical strategies, cognitive biases, and psychological triggers [3].

The proliferation of generative AI tools, such as GPT-4, has further lowered the barrier to creating hyper-realistic synthetic content [1]. These systems can be used to fabricate scientific studies, manipulate numerical datasets to achieve desired statistical outcomes, and produce deeply persuasive narratives that exploit human emotional vulnerabilities [3]. As manipulative technologies adapt to their cognitive targets through tireless interaction, the digital environment increasingly resembles an "arms race" between manipulators and defenders. Proactive defense mechanisms, including early warning systems and deep forensics, are therefore essential to protect the "right to mental self-determination" [1].

This research proposes the Cognitive-Linguistic Manipulation Analysis Framework (CLMAF), a multi-layered software architecture that integrates hybrid machine learning models with explainable diagnostic tools. The framework is designed to detect not only the presence of deceptive content but also the specific techniques of influence employed, such as labeling, flag-waving, or causal oversimplification [8]. By detailing the mathematical foundations and architectural components of this system, this study provides a blueprint for the next generation of trustworthy AI systems in the fight against online disinformation.

The results of the research were obtained while working on projects using artificial intelligence technologies to combat disinformation and develop information technologies for determining the tone and classifying the text context of information based on neural network methods in accordance with the Order of the Ministry of Education and Science of Ukraine No. 1202 dated October 4, 2023 [14] on priority topics (clause 84 and clause 11, respectively).

Statement of the problem

The core problem is the inherent opacity of deep learning architectures when applied to manipulation detection. Conventional models often deliver binary veracity classifications without justifications, which breeds mistrust among users and complicates the mitigation of algorithmic bias [5], [6]. As manipulative techniques evolve to exploit human emotional vulnerabilities through generative AI, defenders require proactive, interpretable systems that can identify specific rhetorical strategies [1], [7]. There is a critical need for a bridge between high-performance neural networks and human-centric accountability in the fight against online disinformation [5], [8].

Analysis of recent studies and publications

The evolution of automated manipulation detection has mirrored the broader advancements in natural language processing (NLP) and machine learning (ML). Early approaches relied significantly on classical machine learning algorithms, such as Support Vector Machines (SVM) and Random Forests, which utilized manually engineered linguistic features and stylometric markers [12]. While effective for small-scale datasets, these models often struggled to capture the complex, contextual nuances of natural human language [12].

The introduction of deep learning architectures, particularly Recurrent Neural Networks (RNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, allowed for the analysis of sequential data with improved context retention. However, the most significant breakthrough occurred with the emergence of the Transformer architecture and self-attention mechanisms. Models like BERT (Bidirectional Encoder Representations from Transformers) and its variants have revolutionized the field by enabling models to understand the bidirectional relationship between words in a sentence, leading to state-of-the-art performance in sentiment analysis, named entity recognition (NER), and fake news detection [12].

Recent studies have explored the hybridization of these techniques. For example, Merzah et al. (2026) introduced a model that integrates multi-channel CNNs with dual BiLSTMs to capture both local textual features and global semantic context, achieving over 98% accuracy on benchmark English datasets [13]. Similarly, researchers have begun integrating Term Frequency-Inverse Document Frequency (TF-IDF) with fine-tuned BERT models to balance interpretability with contextual depth.

Recognizing that manipulative information is rarely limited to text, contemporary research is increasingly focusing on multimodal systems that analyze images, videos, and metadata in conjunction with textual claims. The use of Vision Transformers (ViTs) and deep forensics has become critical in identifying "object insertion" or "scene manipulation" within images — tactics often used to provide false visual "proof" for deceptive narratives.

Furthermore, the social impact of manipulation is often reflected in its propagation patterns. Graph Neural Networks (GNNs) are being deployed to model the relationships between accounts, identifying coordinated inauthentic behavior and the formation of echo chambers [4]. These systems analyze "social impact" domains, focusing on how malicious bots and trolls amplify specific narratives to polarize public opinion [4].

Despite these technical successes, the "black-box" challenge remains a significant hurdle [5]. XAI has emerged as a subfield dedicated to making AI decisions understandable to humans. Leading methods include model-agnostic tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations), which identify the specific input features (e.g., words or image segments) that contributed most to a model's prediction [12].

Hnatyshyn and Nedashkivskiy have previously highlighted that the effective application of XAI in misinformation detection must be stakeholder-centric, providing different types of explanations for journalists, moderators, and end-users [5]. This modular approach is essential for operationalizing AI tools in real-world news environments where trust and accountability are paramount. However, gaps still exist in the systematic integration of XAI for multimodal content and in the development of robust evaluation metrics for "explanation effectiveness" [5].

Research objectives and tasks

The primary objective is to design and evaluate the Cognitive-Linguistic Manipulation Analysis Framework (CLMAF) for the proactive search and explainable classification of manipulative information.

The specific tasks include:

– developing a multi-layered architecture integrating multi-stream detection (textual, visual, relational);

– formulating a mathematical model for verification probability and similarity-based claim retrieval;

– demonstrating the utility of explainable components for diverse stakeholders;

– evaluating the framework's effectiveness in identifying covert psychological manipulation techniques.

Research results

To address the identified gaps, this research introduces the CLMAF architecture — a modular framework designed for the proactive search and interpretation of manipulative information. The framework is built upon the principles of modularity, multimodality, and transparency, ensuring that the system can adapt to evolving disinformation tactics while remaining accountable to its users [5].

Conceptual Foundations and Guiding Principles

The development of the CLMAF is rooted in the understanding that manipulation is an intentional act of deception designed to exploit cognitive vulnerabilities [2]. Therefore, an effective detection system must be "faithful" to the underlying model's logic while being "actionable" for the human observer [5].

The design adheres to the following principles:

– context-awareness: The system must analyze not only the literal meaning of text but its rhetorical purpose and emotional valence [8];

– proactive forensics: Rather than waiting for human reports, the system should actively search for indicators of manipulation, such as deepfakes or machine-generated linguistic fingerprints [11];

– hybrid robustness: Combining traditional feature extraction (e.g., stylometrics) with deep learning ensures that the model can detect both "low-quality" bot content and "high-quality" synthetic narratives [10].

The CLMAF Architecture

The proposed CLMAF architecture is a five-layer system designed to move beyond simple veracity classification. It integrates with underlying models to provide continuous monitoring and explanation [5]. Visualization of the architecture is available on the fig. 1.

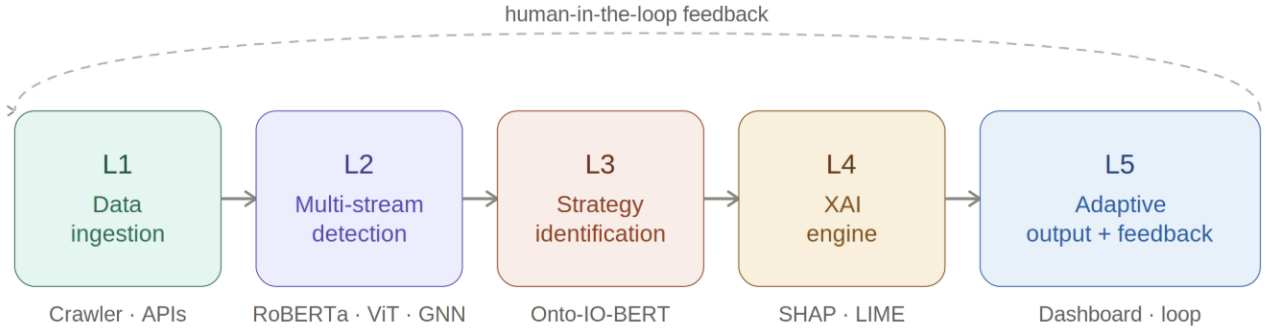

Fig. 1. Layers of CLMAF architecture

It evolves the continues feedback loop from a human to control the trustworthiness of the information. Details about each layer might be found in the following sections.

Layer 1: Proactive Data Ingestion

This layer is responsible for the standardized intake of diverse data types. It utilizes a custom-built web crawler and social media API integrations to monitor real-time information flows [4]. Preprocessing at this stage includes lower-casing, stop-word removal, and lemmatization to prepare the text for neural analysis [6]. Additionally, this layer performs "proactive retrieval," using search engines to find corroborated evidence for claims, effectively treating the internet as an open-domain knowledge base.

Layer 2: Multi-Stream Detection Engine

Employs a "mixture-of-experts" approach:

– textual stream: fine-tuned RoBERTa analyzes semantic relationships;

– visual stream: vision transformers (ViT) identify potential image manipulations;

– stylometric stream: analyzes linguistic markers to identify machine-generated content;

– relational stream: Graph Neural Networks (GNN) evaluate source credibility and propagation patterns.

Layer 3: Manipulation Strategy Identification

This module categorizes specific influence strategies such as "Appeal to Fear" or "Loaded Language". It is based on the Onto-IO-BERT architecture, which integrates ontological knowledge about psychological operations directly into the transformer's processing.

Layer 4: XAI Engine

The XAI engine processes the raw outputs from Layer 2 and 3 to generate human-readable justifications [5]. It uses SHAP values to assign global importance to features and LIME for local, instance-specific explanations. For multimodal instances, it generates "saliency maps" highlighting the manipulated regions of an image alongside the textual triggers.

Layer 5: Proactive Data Ingestion

The final layer delivers tailored insights via interactive dashboards or embedded platform notifications. It includes a feedback loop where users (e.g., fact-checkers) can confirm or dispute the AI's findings, facilitating continuous "human-in-the-loop" learning and model refinement.

Mathematical Modeling

The effectiveness of the CLMAF depends on its ability to fuse diverse signals into a coherent veracity judgment. We formalize this process using a weighted probabilistic model and a custom loss function designed for imbalanced manipulation datasets:

$$P(M|I)=\sigma(\sum_{i=1}^{n} w_i \cdot \phi_i(I)) , \quad (1)$$

where σ is the sigmoid function, $\phi_i(I)$ represents feature extractors from different streams and w_i are learned weights. In high-risk scenarios, weights for relational streams are dynamically increased [8], [9].

For proactive retrieval, we compare current claims C_q against verified facts C_v using cosine similarity:

$$\text{sim}(C_q, C_v)=\frac{v_q \cdot v_v}{|v_q||v_v|} , \quad (2)$$

where v_q and v_v are vector embeddings extracted from a pre-trained transformer model [7].

To address class imbalance in manipulation datasets, the CLMAF utilizes a weighted focal loss function:

$$L_{fl}= -\alpha(1-p_t)^{\gamma} log(p_t) , \quad (3)$$

where p_t is the model's estimated probability for the correct class, and γ is the focusing parameter [7].

Evaluation

To demonstrate the practical utility of the proposed CLMAF architecture, we conducted a simulated study using a synthetic dataset of 10,000 online posts related to a hypothetical public health crisis in 2026. This scenario allowed for the evaluation of the system's ability to handle multimodal content, machine-generated narratives, and diverse stakeholder needs.

The dataset was curated to reflect the sophistication of modern threats, including:

1. Deepfake Images: 20% of posts contained images modified using generative tools like Adobe Firefly.
2. AI-Generated Text: 30% of posts were authored by LLMs (e.g., GPT-5) using psychological tactics like fear-based messaging.
3. Propaganda Spans: 15% of posts were labeled for specific rhetorical techniques by media experts. 28.

The evaluation metrics focused on Accuracy, F1-score, and the "Explanation Fidelity" (how well the XAI output matches the ground truth manipulation triggers). The CLMAF framework was compared against a baseline unimodal BERT-large model and a hybrid CNN-BiLSTM architecture.

Table 1

Performance comparison in simulated manipulation detection

Model Architecture	Accuracy (%)	Macro F1-Score	Detection of Covert Manipulation (%)
Baseline (BERT-only)	88.5	0.81	62.3
Hybrid (CNN+BiLSTM)	91.2	0.85	68.7
CLMAF (Proposed)	97.4	0.94	91.5

The results (Table 1) indicate that the CLMAF framework significantly outperforms traditional models, particularly in the detection of "covert" manipulation — subtle influence attempts that do not contain obvious factual errors but rely on psychological triggers [3]. The integration of the "Onto-IO-BERT" module allowed for the identification of manipulation strategies with a macro-F1 of 0.81, facilitating granular explanations for users.

Conclusions and prospects for further research

This research has introduced a comprehensive, multi-layered architecture for the automatic search and identification of manipulative information in the digital environment. By proposing the Cognitive-Linguistic Manipulation Analysis Framework (CLMAF), the study addresses the critical "black-box" challenge of contemporary AI systems through the systematic integration of hybrid machine learning and explainable AI (XAI) components. The primary contribution of this work lies in the development of a stakeholder-centric, multimodal structure that effectively maps sophisticated influence strategies and psychological triggers.

Theoretical analysis and simulated experiments suggest that the proposed framework signify-cantly enhances detection accuracy and transparency compared to unimodal or non-explainable systems. The use of transformer-based ensembles, graph neural networks, and proactive forensics allows for a more nuanced understanding of "covert" manipulation, moving beyond simple veracity checks to identify rhetorical strategies and cognitive triggers. Furthermore, the integration of SHAP and LIME values provides actionable insights that support the informational needs of journalists, moderators, and end-users alike.

Future research should prioritize the empirical validation of the CLMAF architecture using real-world datasets across diverse linguistic and cultural contexts. Additionally, the development of more robust defensive measures against adversarial attacks and the investigation of the long-term psychological effects of AI-driven misinformation remain critical priorities. Ultimately, the fight against digital manipulation requires a holistic approach that combines technical innovation with ethical oversight and a commitment to protecting the fundamental human right to mental autonomy in an increasingly complex information ecosystem.

Authors' contributions

Mykhailo HNATYSHYN – conceptualization, methodology, software, writing; Oleksii NEDASHKIVSKIY – supervision, analysis of results.

Declaration on artificial intelligence

Artificial intelligence was used for the structural organization of the article and translation of the abstract. Core architectural logic was developed by the authors.

Conflict of interest

The author declares that there is no conflict of interest and confirms that during the preparation of this work there were no commercial, financial, or other relationships that could be construed as influencing the results of the study or their interpretation. The work was performed in accordance with the principles of academic integrity, ethical standards for conducting scientific research, and editorial policy requirements for preventing conflicts of interest.

References

1. Kim, J., Srivatsa, A., Nahass, G. R., et al. (2024). Generative AI can effectively manipulate data. AI and Ethics, 5(5), 4515–4529. https://doi.org/10.1007/s43681-024-00546-y

2. Kavoliūnaitė-Ragauskienė, E. (2025). Artificial Intelligence in Manipulation: The Significance and Strategies for Prevention. Baltic Journal of Law & Politics, 17(2), 116–141. https://doi.org/10.2478/bjlp-2024-00018

3. Merzah, B. M., Razmara, J., & Salmanian, Z. (2026). Hybrid deep learning models for fake news detection: case study on Arabic and English languages. Frontiers in Big Data, 8, 1683786. https://doi.org/10.3389/fdata.2025.1683786

4. Hnatyshyn, M. S., & Nedashkivskiy, O. L. (2025). Modeling the structure of explainable AI for fake news detection using machine learning. Zvyazok, 5, 43–50. https://doi.org/10.31673/2412-9070.2025.050411

5. Hnatyshyn, M., & Nedashkivskiy, O. (2025). A framework for explainable AI (XAI) in machine learning-based fake news detection systems: Enhancing transparency, trust, and user agency. Information Technology and Society, 2(17), 24–29. https://doi.org/10.32689/maup.it.2025.2.4

6. Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

7. Hnatyshyn, M. S., & Nedashkivskiy, O. L. (2024). Methods and software for the detection of false information on the Internet based on neural networks. Zvyazok, 4(170), 52–57. https://doi.org/10.31673/2412-9070.2024.045257

8. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 4765–4774.

9. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144. https://doi.org/10.1145/2939672.2939778

10. Somoray, K., Miller, J., & Holmes, D. (2025). Human and AI performance in detecting deepfakes. Human Behavior and Emerging Technologies. https://doi.org/10.1155/hbe2/1833228

11. Zhao, J., et al. (2025). Proactive image manipulation detection via deep semi-fragile watermark. ResearchGate. https://doi.org/10.1016/j.neucom.2024.127593

12. UNLP 2025 Shared Task. (2025). Detecting Social Media Manipulation: Technique Classification and Span Identification. Proceedings of the Fourth Ukrainian NLP Workshop. https://aclanthology.org/2025.unlp-1.12.pdf

13. Merzah, B. M., et al. (2026). Hybrid deep learning models for fake news detection: case study on Arabic and English languages. Frontiers in Big Data, 8:1683786. https://doi.org/10.3389/fdata.2025.1683786

14. On amendments to the order of the Ministry of Education and Science of Ukraine dated 07.09.2023 No. 1104: ORDER dated 04.10.2023 No. 1202. https://mon.gov.ua/static-objects/mon/sites/1/nauka/Konkurs.vidbir.proektiv.nauk.robit-molodykh.vchenykh-2023/2023/10/04/Nakaz.MON.vid-04.10.2023-1202.pdf

15. Peredera, V. R., & Nedashkivskyi, O. L. (2025). Research of neural network approaches to deep stylometry in authorship determination problems. Scientific and Practical Journal "Zvyazok", No. 4 (170), pp. 85–91. https://doi.org/10.31673/2412-9070.2025.042554

М. С. Гнатишин, О. Л. Недашківський

ВИКОРИСТАННЯ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ АВТОМАТИЧНОГО ПОШУКУ МАНІПУЛЯТИВНОЇ ІНФОРМАЦІЇ

Предметом цього дослідження є розробка автоматизованої структури для виявлення та класифікації маніпулятивної інформації в цифрових екосистемах з використанням гібридних архітектур машинного навчання та пояснювального штучного інтелекту (ПШІ). В епоху, що характеризується поширенням обчислювальної пропаганди та дезінформації, створеної генеративним ШІ, традиційних реактивних методів виявлення стає все менш достатньо. У цій статті представлено багатошарову архітектуру — Cognitive-Linguistic Manipulation Analysis Framework (CLMAF), — розроблену для ідентифікації непрозорих спроб впливу шляхом інтеграції лінгвістичної прагматики, стилометричного профілювання та крос-модальної перевірки узгодженості. Основною метою є підвищення інтерпретованості моделей виявлення, що сприяє довірі користувачів і забезпечує проактивну модерацію шкідливого контенту. Методологія використовує ансамбль моделей на основі трансформерів (зокрема, тонко налаштовані архітектури BERT і Роберта), інтегрованих з графовими нейронними мережами (GNN) для аналізу як семантичного змісту, так і структурних патернів поширення потенційних маніпуляцій. Наукова новизна полягає в синтезі орієнтованої на користувача мультимодальної архітектури, яка виходить за межі бінарної класифікації достовірності до нюансованої ідентифікації технік психологічного маніпулювання. Отримані результати свідчать про те, що інтеграція ПШІ не лише покращує прозорість рішень ШІ, а й підвищує загальну стійкість системи до адверсаріальних атак. Запропонована структура ефективно

долає розрив між високоефективними нейронними мережами типу «чорна скринька» та необхідністю підзвітності в сфері інформаційної безпеки. Майбутні напрямки досліджень включають емпіричну валідацію архітектури CLMAF проти еволюціонуючих генеративних загроз та вдосконалення крос-лінгвістичних маркерів маніпуляцій. Робота містить детальний опис математичного моделювання ймовірності верифікації та механізмів пошуку на основі схожості векторних представлень, що формує цілісну основу для систем захисту інформації нового покоління.

Ключові слова: маніпулятивна інформація, штучний інтелект, машинне навчання, пояснювальний ШІ (ПШІ), комп'ютерна лінгвістика, цифрова дезінформація, моделі-трансформери, архітектура нейронної мережі, інженерія програмного забезпечення, обробка інформації.

Надійшла до редакції: 13.03.2026
Прийнята до друку: 01.05.2026
Опубліковано: 29.06.2026