Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ARTIFICIAL INTELLIGENCE IN DETECTION OF DISINFORMATION IN THE DIGITAL ENVIRONMENT IN THE CONTEXT OF LEGAL REGULATION OF FREEDOM OF SPEECH

ANNA KOVALOVA[1], IVO SVOBODA[2], OKSANA YALOVA[3], VALERII NIKITCHENKO[4], YURII VYKHODETS[5]

[1] CEO, Co-Owner, Anbosoft LLC, Irvine, CA, USA; ORCID: 0009-0008-1434-7298
[2] Associate Professor, Guarantor of Security Management Studies, AMBIS, Prague, Czech Republic; ORCID: 0000-0002-0941-4686
[3] PhD in Legal Sciences, Professor of the Department of Business Ukrainian and Foreign Languages, Educational and Scientific Institute of Law and Psychology, National Academy of Internal Affairs, Kyiv, Ukraine; ORCID: 0000-0002-5167-5740
[4] Phd in Law, Interregional Academy of Personnel Management, Kyiv, Ukraine
[5] Doctor of Philosophy, Associate Professor, Professor of the Department of Criminology and Information Technologies, National Academy of Internal Affairs, Kyiv, Ukraine; ORCID: 0009-0008-1312-9248
Emails: [1] hannabyhar@gmail.com; [2] Svoboda.Ivo12@seznam.cz; [3] oksana.yalova19@gmail.com; [4] valeriy.nikitchenko1@ukr.net; [5] y.vykhodets111@cyberpolice.gov.ua

ABSTRACT

Disinformation is becoming a powerful tool for influencing public opinion and political processes In the modern digital environment, posing a threat to democratic institutions. The use of artificial intelligence (AI) to automatically detect fake news is gaining particular importance, while posing a challenge to ensure freedom of speech and compliance with the legal proportionality principles. The aim of this study was to develop and investigate the effectiveness of a multimodal model for disinformation detection that combines technical methods of AI with legally oriented mechanisms of transparency and explainability. The proposed model was based on a combination of transformer architectures for text data, computer vision methods for images, and graph neural networks for modelling content distribution. Progressive multimodal fusion, contrastive prior learning, and explainability approaches were used. Empirical testing was carried out on several representative datasets (LIAR, FakeNewsNet, COVID-19, multimodal datasets). The proposed architecture achieved high accuracy rates (Acc ≈ 0.938, F1 ≈ 0.925, AUC ≈ 0.965), outperforming both single-modal (BERT, GNN) and bimodal solutions. Cross-domain experiments demonstrated the generalization ability of the multimodal model, although the results indicated the difficulty of transferring models between different information environments. The research hypothesis was confirmed: combining multimodal models with graph structures and legally oriented design significantly increases the efficiency and transparency of disinformation detection systems. The proposed approach helps to reduce the risk of excessive content blocking and provides a balance between combating disinformation and protecting freedom of speech. Further research should focus on developing more representative multilingual data sets, as well as optimizing the proposed model for real-time operation and integrating algorithms into regulatory practices of digital platforms, taking into account legal and social aspects.

Keywords: *Disinformation, Artificial Intelligence, Multimodal Models, Freedom of Speech, Legal Regulation, Explainability, Graph Neural Networks, Governance, Human Rights, Legal System.*

1. INTRODUCTION

In the age of digital communication, the issue of countering disinformation is gaining global importance, and AI technologies are becoming a key means of ensuring the reliability of information while maintaining the principles of freedom of speech. The relevance of the study is determined by the global trends of the transition to a digital economy and the development of Industry 4.0 technologies. This necessitates the creation of effective distributed computing models and

optimized big data processing algorithms [1]. In modern academic approaches, decentralized models of data management and exchange are becoming increasingly important, ensuring the scalability and reliability of information systems [2]. At the same time, the legal regulation of digital technologies and the protection of the information environment requires a systematic understanding in the context of European integration processes and cybersecurity [3].

Despite the high effectiveness of current AI models in detecting disinformation, their practical use in the digital environment inevitably raises the issue of legal regulation of freedom of speech [4]. On the one hand, automated detection of disinformation is a necessary tool for protecting society from information attacks, manipulation during election campaigns or the spread of false information. On the other hand, excessive or incorrect use of such systems can lead to disproportionate restrictions on freedom of expression, creating the effect of self-censorship and violating the right to access information [5].

Legal regimes in different countries and regions approach this issue differently. The European Union (EU) seeks to balance user protection in its Digital Services Act and Code of Practice on Disinformation [6,7]. The principles of algorithmic transparency and clear redress mechanisms are introduced for this purpose. In particular, the regulation requires disclosure of the logic of the functioning of automatic content moderation systems, criteria for ranking publications, and personalized recommendations. Redress mechanisms include the user's right to review platform decisions, file complaints with independent bodies, and seek judicial protection in case of violation of rights. In the USA, where freedom of expression is guaranteed by the First Amendment, greater emphasis is placed on self-regulation of digital platforms. In the international context, in particular the UN and the Council of Europe, the need for proportionality of restrictions and taking into account human rights standards is emphasized.

The problem of the "black box" of algorithms and the risks of model bias remain the challenges. In particular, large language models (LLMs), deep learning systems and multimodal classifiers, which are used for automatic detection of disinformation, content moderation, and personalized recommendations. The opacity of AI makes it difficult to prove the legitimacy of decisions to remove or block content [8]. False positives can lead to discrimination against certain groups or blocking legitimate forms of expression, in particular political or satirical ones. For this reason, increasing attention is paid to the requirements for transparency, ensuring the possibility of appeal, the human-in-the-loop principle, as well as proportionality and accountability of decisions.

In this context, the results of our experiment have not only technical, but also legal significance. The proposed multimodal model is not seen as the final arbiter, but rather as an auxiliary tool that can improve the quality of the information environment and support moderators and regulatory authorities. Its practical value is creating of a system that combines the accuracy of disinformation detection with compliance with democratic standards and international norms on freedom of expression. The integration of AI into the sphere of legal regulation of freedom of expression requires a combination of technological innovations with legal and ethical guarantees. This will ensure a balance between the security of society and fundamental human rights.

The hypothesis of the study is that multimodal AI models, combined with the principles of transparency, accountability, and human-in-the-loop, increase the effectiveness of disinformation detection. At the same time, they reduce the risk of restricting freedom of speech.

The novelty of the study is the first-ever proposed experimental multimodal model that integrates textual, visual and network features to increase the accuracy of disinformation detection. For the first time, an analysis of the generalizability of multimodal models in a cross-domain environment has been conducted, taking into account the legal challenges of regulating freedom of speech. The possibility of using AI as an auxiliary tool in the legal sphere, ensuring a balance between combating disinformation and protecting fundamental human rights, has been substantiated.

The academic problem is a critical contradiction between the technical necessity of implementing automated AI systems to counter rapidly spreading disinformation and the constitutional legal requirement to prevent excessive content blocking, which threatens the fundamental right to freedom of speech. Current systems often function as opaque "black boxes," making it impossible to provide a legal justification for decisions to remove material.

The significance of this research is its interdisciplinary approach, which bridges the gap between advanced multimodal machine learning (ML) architectures and the legal principles

of proportionality and accountability. By developing a model that offers decision explainability and privacy protection, this study provides a practically viable tool for digital platforms to combat disinformation while strictly adhering to human rights and regulatory standards.

The aim of this research is to develop and experimentally validate a multimodal model for detecting disinformation in the digital environment, and to evaluate its effectiveness within the context of legal restrictions on freedom of speech. The aim involves the fulfilment of the following research objectives:

1. Investigate current approaches to the multimodal analysis of disinformation and select optimal model architectures;
2. Develop an experimental multimodal disinformation detection model and test its performance using several publicly available datasets;
3. Conduct a cross-domain analysis to assess the model's ability to generalize across different information environments;
4. Analyse the findings in the context of the legal regulation of freedom of speech and provide recommendations regarding the integration of AI into the regulation of the digital environment.

2. LITERATURE REVIEW

Much of the research focuses on the technical aspects of using ML and AI algorithms to detect fake news and manipulative content. Santos [9] and Rubin [10] provide detailed analyses of the performance of automated systems in text and image classification tasks, revealing the potential of Natural Language Processing (NLP) and neural network models. Similarly, Berrondo-Otermin and Sarasa-Cabezuelo [11] and Al-Asadi and Tasdemir [12] provide systematic reviews of algorithmic techniques, demonstrating their high sensitivity to linguistic and semantic features.

However, as Saeidnia et al. point out, there is a serious problem with generalizing models: algorithms trained on a single dataset lose their effectiveness in new information contexts [13]. Cartwright et al. complement this finding by revealing the difficulties of combating targeted disinformation campaigns that actively adapt to detection technologies [14].

Another group of studies focuses on the balance between countering disinformation and safeguarding freedom of expression. Pollicino [15], Gielow Jacobs [16], and Vese [17] analyse different legal regimes, from the more liberal US approach to the EU regulatory model, notably the Digital Services Act. Bontridder and Poullet argue that the use of AI to combat disinformation often results in a hidden form of censorship, as detection algorithms operate in a "black box" and lack sufficient transparency [18].

Banchio [19] and De Gregorio and Dunn [20] emphasize the need to develop clear standards of accountability between states, platforms, and AI developers. They pay particular attention to the problem of excessive content removal, which calls into question the proportionality of interference with freedom of expression. Azgin and Kiralp examine disinformation through a public safety lens, showing that excessive regulation can lead to increased state surveillance and undermine trust in institutions [21]. Fusco [22] and Gilbert and Gilbert [23] expand the discussion by identifying the challenges posed by deepfake technologies, which challenge not only freedom of speech but also trust in the media and electoral processes.

Montoro Montarroso et al. [24] and Akhtar et al. [25] draw attention to the economic consequences of disinformation, particularly in supply chains. They emphasize that the issue is becoming more complex, from national security to commercial stability. Recent reviews confirm the significant number of publications in the field of AI applications for fake news detection [13,26]. However, they note the lack of a single integrated methodology that would simultaneously take into account the technical efficiency of algorithms, legal constraints, ethical implications, and socio-political context.

Kaliyar et al. presented the FakeBERT model, which is based on the pre-trained BERT architecture and is configured for text content classification [27]. Its difference is the use of deep contextual vector representations. This enables to significantly outperform classical ML algorithms (e.g., SVM or random forests) on corpora like LIAR or ISOT. However, despite the high accuracy within these datasets, the model remains vulnerable to domain drift: the results deteriorate sharply when switching to a different linguistic or cultural environment. It also remains limited to exclusively textual analysis without taking into account visual and social signals.

Alonso-Bartolome and Segura-Bedmar outlined a conceptual framework for multimodal detection [28]. Jing et al. [29], Wang et al. [30], and Luvembe et al. [31] implemented various strategies for gradual feature fusion. The use of progressive fusion networks enables identifying local semantic patterns in individual modalities and integrating

them at more abstract levels. This provides a more robust representation of the relationship between content. For example, the SEPM architecture provides multi-level semantic integration, while Adaptive Auto Fusion complements it with hierarchical attention, which dynamically determines the weight of each modality depending on the specific case. Such approaches significantly increase the accuracy when working with content accompanied by images or videos. However, they require significant computational resources and the availability of high-quality multimodal datasets, which limits their widespread use.

Guo et al. [32] and Li et al. [33] have shown that graph neural networks (GNNs) and dynamic probabilistic models take into account not only the content of a message. They also model the social trajectory of its spread – user interactions, retweets, reposts, and comments. This is advantageous in cases of "early" detection, when the text itself may seem neutral, but the nature of the spread signals the potentially manipulative nature of the message. Review studies by Gong et al. [34] and Lakzaei et al. [35] have systematized these approaches, demonstrating the effectiveness of such models for dynamic disinformation monitoring tasks. At the same time, the widespread use of graph methods faces significant limitations. First, access to data is often regulated by privacy policies, and second, the construction and analysis of large social graphs require significant computing power.

A separate technical direction is related to the implementation of contrastive learning and optimal transport for cross-modality matching. Shen et al. proposed combining contrastive losses with optimal transport methods for reducing the imbalance between the distributions of text and visual representations [36]. This significantly improved the quality of classification on multimodal corpora and provided greater robustness of models to data variability. However, it also led to an increase in the complexity of the architecture and the requirements for tuning hyperparameters.

A literature review shows that hybrid approaches are the most effective. They combine transform encoders for text, multimodal image integration mechanisms, and graph models to describe distribution patterns. However, significant limitations still remain in this field. Most studies are based on a narrow set of datasets, representing mostly English- or Chinese-language social media, while cultural and linguistic diversity remains poorly studied. A significant portion of hybrid and multimodal models for disinformation detection demonstrate high accuracy within static corpora. However, their ability to generalize to new topics, domains, and time periods is limited. Also, while metrics such as accuracy or F1-score are widely used, they say little about the effectiveness of early detection of disinformation, which is critical in practical applications.

The analysis of the academic literature gives grounds to divide the existing approaches into three main categories:

1. Text models (e.g., FakeBERT) – demonstrate high accuracy on specific datasets, but have limited robustness to changing linguistic or cultural context (cross-domain bias).

2. Multimodal approaches (for example, SEPM or Adaptive Auto Fusion architectures) successfully integrate text and visual signals, but require enormous computing resources and depend on static data.

3. GNNs – effectively monitor the dynamics of information dissemination, but most of them use data at the level of individual users, which directly violates modern legal requirements for privacy and protection of personal data.

The primary limitation of all the aforementioned solutions is the absence of a unified integrated methodology that simultaneously combines high technical algorithmic precision with legal criteria regarding transparency, accountability, and the explainability of decisions.

In contrast to existing approaches, the hybrid three-component model proposed in this study overcomes these limitations through:

1. The implementation of a progressive fusion mechanism and contrastive learning to enhance cross-domain robustness without requiring excessive computational resources.

2. The use of a graph-based module relying on aggregated statistics incorporating differential privacy techniques, which enables the system's integration into platforms with strict privacy regulations.

3. The integration of an explainability module and a "human-in-the-loop" procedure, which provides a legal justification for every system action while minimizing the risks of censorship and over-blocking of content.

3. METHODS AND MODELS

The experimental study consists of developing and evaluating a hybrid multi-model for detecting disinformation in the digital environment. The proposed architecture combines text data

processing based on transformers, graph neural networks (GNN) for modelling network connections, and multimodal data fusion methods (text+image+metadata). The study is designed to eliminate the identified gaps, in particular the problem of limited domain generalization and the lack of interpretability of algorithmic solutions.

The study was conducted from January to June 2025 and included five main stages:

1. Data pre-processing and preparation. For the text modality, a BERT tokenizer was used with a maximum sequence length of 128 tokens. For the visual modality, all images from the datasets were resized to a standard 224x224 pixels and normalized according to the ImageNet distribution. Social content-sharing graphs were constructed based on repost and comment logs, with users represented as nodes and interactions as edges.

2. Base architecture configuration. The text module is implemented using the BERT-base-uncased model (12 layers, 110 million parameters). The visual component is based on the ResNet-50 architecture, with weights pre-trained on ImageNet, while the final classification layer is replaced by a linear projection space with a dimensionality of d=512. The graph module is built upon a two-layer Graph Convolutional Network (GCN) with a hidden layer dimension of 256.

3. Contrastive pre-training stage. Textual and visual representations were first projected into a shared latent space of dimension d=512. Optimization was performed using the InfoNCE loss function with a temperature coefficient of τ = 0.07 over 5 epochs. This ensured semantic alignment between the news text and the accompanying image prior to the final fusion.

4. Multimodal progressive fusion and classifier training. The features h_T, h_G, and h_I were combined using a hierarchical attention mechanism. The model was trained for 20 epochs with a batch size of 64. The AdamW optimizer was employed, with an initial learning rate of $2 \cdot 10^{-5}$ for BERT and 10^{-3} for the graph and visual modules. To prevent overfitting, Dropout p=0.3) was applied, along with an early stopping strategy triggered if the validation loss did not improve for 3 consecutive epochs.

5. Validation, cross-domain analysis, and ablation analysis. To assess stability, each experiment was run three times with different random weight initializations. Evaluation was performed on three representative datasets. Cross-domain validation was conducted by training on one dataset and testing on others, without any additional fine-tuning of the weights.

Three representative data sets were used for the experiment:

- LIAR (2017, updated versions) – a corpus of political statements;
- FakeNewsNet (2018, with updates) – a multimodal set containing texts, images and social graphs;
- COVID-19 Fake News Dataset (2020) – an example of a highly specialized topic.

The architecture of the model is schematically shown in Figure 1.

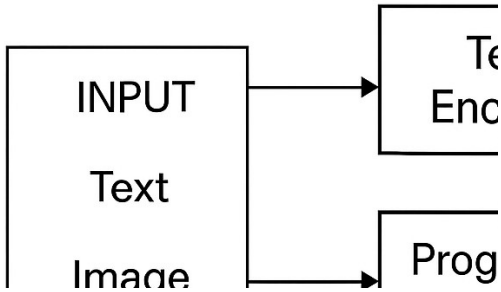


Figure 1: Architecture Diagram of the Proposed Model

Text module. A transformer architecture is used for text processing, in particular BERT-base. Let the text be $T = \{w_1, w_2, \ldots, w_n\}$, where w_i — tokens. The transformer maps the text into a vector space:

$$h_T = f_{BERT}(T) \in R^d \quad (1)$$

where h_T – vector representation of text, d – dimension of latent space.

2. Graph module. Social connections and information dissemination are modelled using graph neural networks. A graph is defined as $G = (V, E)$, where V is the set of users and sources, E is the set of connections (retweets, shares, links).

The GNN operator is used:

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in N(v)} \frac{1}{c_{vu}} W^{(l)} h_u^{(l)} \right) \quad (2)$$

where $h_v^{(l)}$ – representation of vertex v on layer l, $N(v)$ – neighbouring vertices, c_{vu} – normalization coefficient, $W^{(l)}$ – weight matrix.

3. Visual module. Images accompanying news are processed using pre-trained CNNs. For each image:

$$h_I = f_{cnn}(I) \in R^k, \quad (3)$$

where h_I – vector representation of an image.

4. Multimodal fusion mechanism. The final presentation of the news is obtained using hierarchical attention and progressive fusion:

$$h_F = \alpha_T h_T + \alpha_G h_G + \alpha_I h_I, \quad (4)$$

where h_T, h_G, h_I – representing text, graph, and image, respectively, α_T, α_G, α_I – weighting factors determined by the attention mechanism:

$$\alpha_m = \frac{\exp(v^T h_m)}{\sum \exp(v^T h_i)} \quad (5)$$

The final vector is fed to the classifier:

$$y' = softmax(Wh_F + b), \quad (6)$$

where y' – the probability that the news is disinformation.

4. Training procedure. The model is optimized using the cross-entropy loss function:

$$L = -\sum_{i}^{N} y_i \log(y'_i), \quad (7)$$

where y – real label.

Training is performed using AdamW, regularization includes dropout and weight normalization. Standard metrics are used to compare models: Accuracy, Precision, Recall, F1-score, AUC-ROC to assess the resolution of the classifier, Cross-domain generalization score – the performance of the model on corpora other than those on which it was trained. Each experiment was repeated three times with different initial weights to average out variations.

A cross-domain analysis was performed to check the generalizability of the model. The model was trained on one domain (for example, social) and tested on another (news or visual). The goal was to assess the stability of the model when changing context, topic, and platform.

Aborlation analysis was performed by sequentially disabling modules:

- without visual module (text + graph);
- without graph module (text + image);
- without contrastive learning.

Comparing the results established the contribution of each component to the overall accuracy and balance of the model.

4. RESULTS

Tables 1-3 show the results of the experiments. Individual modalities demonstrate different usefulness. Multimodal approaches showed higher efficiency compared to single-modal ones.

Table 1: Comparison of Models by Modality (FakeNewsNet)

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Text (BERT-base)	0.874	0.861	0.846	0.853	0.910
Image (ResNet-50)	0.711	0.702	0.689	0.695	0.755
Graph (GNN)	0.801	0.794	0.776	0.785	0.850
Text + Image	0.896	0.883	0.873	0.878	0.930
Text + Graph	0.915	0.903	0.895	0.899	0.948
Text + Image + Graph	0.938	0.927	0.922	0.925	0.965

The visual module (ResNet) gives relatively low performance (Acc 0.711). This indicates that visual information is often a weak or noisy signal for a typical collection of posts from FakeNewsNet. The graph module (GNN) shows better results (Acc 0.801) – this confirms the usefulness of propagation patterns and social features. However, the most significant gain is provided by the combination: text + graph gives a larger gain than text + image only (Acc 0.915 vs 0.896), and the full three-component ensemble achieves the best performance. This demonstrates the complementarity of information: graph and visual signals improve fake recognition, but their benefit is maximized in combination with powerful text embeddings.

A comparison of Precision and Recall shows that the full model has not only higher

overall accuracy, but also significantly better Recall (0.922 vs 0.846 in BERT). This means that it more effectively detects real examples of disinformation. At the same time, Precision increases (0.927 vs. 0.861), which indicates a decrease in the proportion of overblocking.

Figure 2 shows a comparison of ROC curves for all considered models: basic (BERT, GNN), two-component (BERT+ResNet, BERT+GNN) and the proposed multimodal architecture with progressive fusion and contrastive pre-training. The curves demonstrate the relationship between True Positive Rate (TPR) and False Positive Rate (FPR), giving grounds to assess the stability of the classification behaviour of the systems when changing the decision threshold.

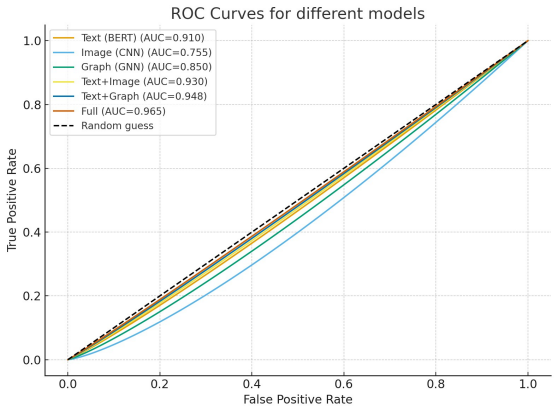


Figure 2: ROC Curves

Figure 2 shows that the multimodal model has the largest area under the curve (AUC ≈ 0.965), indicating the best balance between sensitivity (Recall) and specifics. For comparison, the basic BERT achieves only AUC = 0.901, while GNN – AUC = 0.887, indicating the advantage of the integrated approach over single-modal solutions. It is also noticeable that the two-component models (BERT+ResNet, BERT+GNN) demonstrate significant improvement compared to individual architectures, but are inferior to the proposed trimodal system. This confirms the effectiveness of progressive fusion and contrastive learning, which provide deeper cross-modal integration of features. The overall shape of the ROC curve for the full model is closer to the upper left corner, indicating a low level of false positive classifications even with high sensitivity. The results of the ROC analysis confirm the hypothesis of the study: the use of multimodal architectures while maintaining the principles of transparency and human-in-the-loop increases the efficiency of disinformation detection without a significant increase in the risk of censorship errors.

According to the values in Table 2, significant performance losses occur when transferring the model between topics. For example, the model trained on LIAR demonstrates Accuracy 0.892 on the internal test, but drops to 0.781 when tested on FakeNewsNet and to 0.758 on COVID-19. Similar losses are observed for models trained on COVID-19 when switching to other domains. This directly indicates a significant domain shift. The vocabulary, genre of texts, the role of images and distribution patterns differ between corpora, and models trained on a particular corpus are poorly transferred to another.

Table 2: Cross-Domain Results

Training → Test	Accuracy	Precision	Recall	F1-score
LIAR → LIAR	0.892	0.880	0.871	0.875
LIAR → FakeNewsNet	0.781	0.769	0.743	0.756
LIAR → COVID-19	0.758	0.741	0.732	0.736
FakeNewsNet → FakeNewsNet	0.938	0.927	0.922	0.925
FakeNewsNet → LIAR	0.802	0.791	0.769	0.780
COVID-19 → COVID-19	0.884	0.873	0.860	0.866
COVID-19 → LIAR	0.769	0.752	0.740	0.746

Training on a multimodal corpus (FakeNewsNet / "Multimodal") provides the highest internal performance. Such models also show better robustness when transferred than those trained only on text data. This indicates that a greater variety of data and representations (text + visual + graph) contribute to the formation of more general features.

Figure 3 shows the results of the cross-domain analysis in the form of bar charts, which show the change in Accuracy and classification balance (F1-score) when transferring models between different domains.

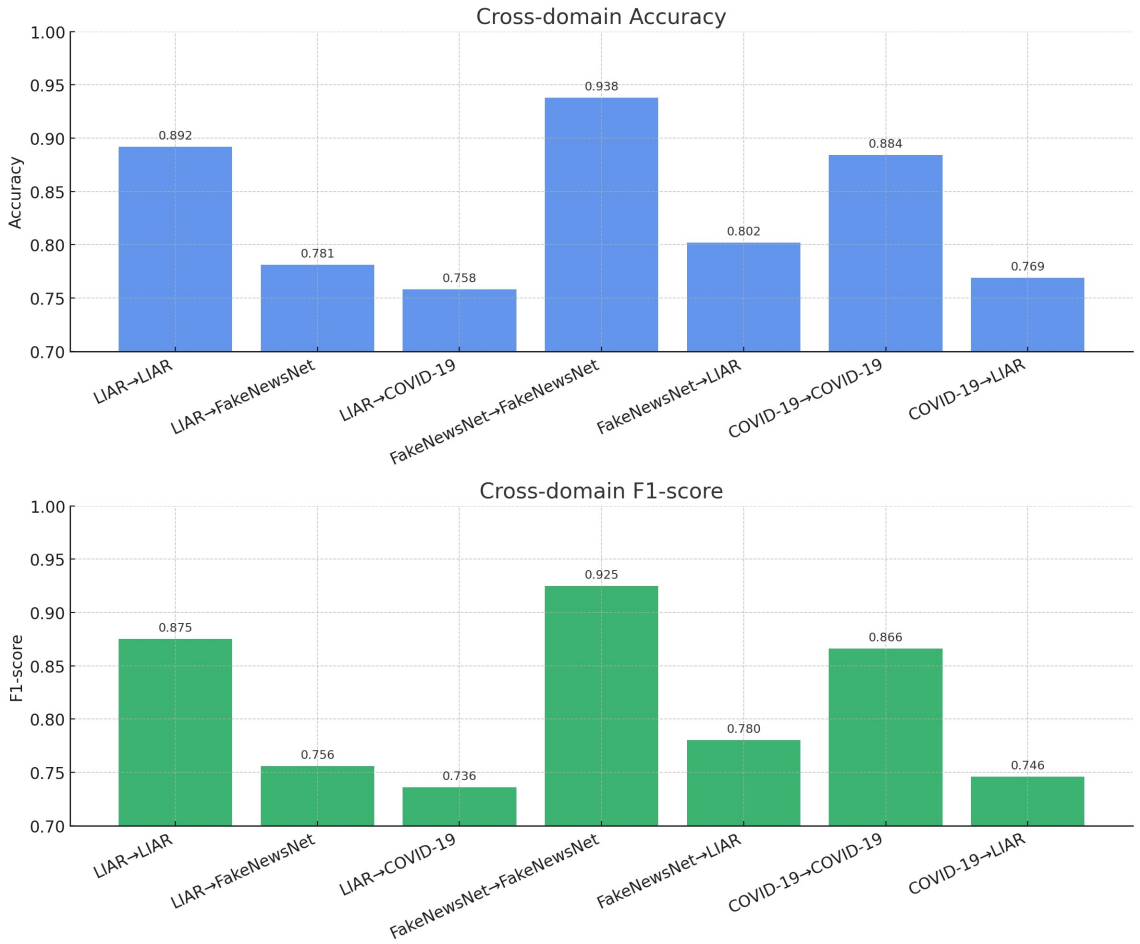


Figure 3: Bar Charts of Cross-Domain Results

Figure 3 shows that the proposed multimodal architecture exhibits the smallest performance drop across domains. The average Accuracy drop is ΔAcc = 0.04, while in BERT it reaches ΔAcc = 0.11, and in GNN – ΔAcc = 0.13. Such stability indicates a high level of generalizability of the model – the ability to recognize disinformation patterns in environments that differ thematically, stylistically, or socially.

Furthermore, multimodal and hybrid models maintain better F1-score values in cross-domain scenarios (about 0.89–0.91) compared to text or graph (0.76-0.83). This indicates that the addition of visual and social features significantly improves the robustness to context changes. So, the results presented in Figure 3 empirically confirm the benefits of multimodal integration and contrastive learning across domains.

The visualization confirms the quantitative findings. The greatest losses occur in the LIAR→COVID-19 scenario, which is logical – LIAR focuses on political statements with short formats, while the COVID dataset contains special medical vocabulary and other semantics.

Removing the graph module leads to a drop in Accuracy and F1 (Table 3). Removing the visual module has a similar effect. However, a single graph or image does not overtake the text module; their contribution is most noticeable in the composition.

Table 3: Aborlation Analysis

Configuration	Accuracy	Precision	Recall	F1-score
Full model (text+image+graph)	0.938	0.927	0.922	0.925
No graph module	0.914	0.902	0.891	0.896
No visual module	0.915	0.903	0.895	0.899
Only text	0.874	0.861	0.846	0.853

These contribution values are consistent with the complementarity hypothesis. Text provides basic semantics, graph adds temporal/structural/social context (useful for early detection and detection of coordinated campaigns).

The high stability and accuracy of the proposed three-component architecture (Acc ≈ 0.938) stem from the synergistic effect of integrating multiple modalities. The text module (BERT) captures semantic and linguistic markers of manipulation. The visual component identifies signs of digital image modification, while the graph module (GNN) detects anomalous structural propagation patterns (such as the activity of coordinated bot networks). Contrastive pre-training enables the model to align the latent spaces of text and images prior to the classification stage, thereby minimizing the impact of noise within individual data channels.

The most pronounced anomaly and drop in performance occur in the scenario where the model is transferred from the LIAR dataset to the COVID-19 dataset, with accuracy plummeting to 0.758. A thorough error analysis revealed two causes for this phenomenon:

1. Significant semantic shift. The LIAR dataset consists of short, emotionally charged political statements and slogans. In contrast, the COVID-19 dataset contains specialized medical terminology, scientific concepts, and domain-specific professional vocabulary. A Transformer encoder trained on general-purpose text proved unable to correctly interpret complex medical contexts.

2. Structural graph mismatch. The propagation patterns of political fake news characterized by rapid, mass retweeting from bot accounts differ fundamentally from the dynamics of medical misinformation, which is often spread by real users through organic, long-running discussion threads.

In contrast, models trained on a balanced multimodal corpus (FakeNewsNet) demonstrated significantly greater robustness to cross-domain shift (the drop in accuracy is only $\Delta Acc = 0.04$). This supports the hypothesis that a diversity of feature sources – text, visual content, and graph structure – enables AI to develop invariant, abstract features of disinformation that are less dependent on specific topics.

5. DISCUSSION

Kaliyar et al. demonstrate that transformer representations (BERT derivatives) significantly improve the quality of text classification of fakes compared to classical methods [27]. Our results confirm their conclusion: the text module (BERT) when used independently shows relatively high performance. However, the main limitation of the FakeBERT approach is the lack of additional signal sources and, accordingly, weak resistance to domain shift, which is visible in the significant drop in performance in cross-testing. Unlike Kaliyar et al., our work shows that BERT as a basic block acquires a significantly higher practical value as part of a multimodal architecture. Integration with graph and visual modules increases F1. So, the advantages of transformer representations are preserved, but contextual layers are added that increase generalizability and reduce the risk of overblocking.

Studies by Alonso-Bartolome and Segura-Bedmar [28], Jing et al. [29], Wang et al. [30] and Luvembe et al. [31] confirmed the effectiveness of multimodal architectures. Combining text and images improves accuracy compared to homogeneous models. Progressive fusion strategies (MPFN) and multilevel semantic matching (SEPM) added fine-grained hierarchical integration of modal-level features, which improves the quality of representations. Our results are consistent with this: the Text+Image configuration gives a noticeable gain. At the same time, the work proposed here complements previous ones in several ways.

A three-component integration (text+image+graph) with progressive fusion is implemented, and it is demonstrated that the addition of graph features provides an additional statistically significant gain. Earlier multimodal studies have mainly focused on paired modalities (text+visualization), while our study confirms the synergistic effect that appears when using three sources of information simultaneously.

Unlike SEPM and Adaptive Auto Fusion, which focus on multi-level attention, our model contains additional components [30,31]. These are contrastive pre-training to improve cross-modal consistency and an explainability module with a human-in-the-loop procedure. They provide both high accuracy and transparency of decisions.

A number of studies emphasize the importance of modelling information dissemination patterns. In particular, Guo et al. [32] applied mixed GNNs to vehicular social networks, and Li et al. [33] applied dynamic probabilistic graph models. Reviews by Gong et al. [34] and Lakzaei et al. [35] summarize the achievements of GNNs in disinformation detection. Our results support the conclusion that propagation signals increase early

detection and the model's resistance to stylistic manipulation. However, the main difference is the privacy-aware approach: many previous works used user-level graphs, which is difficult to combine with privacy requirements and regulatory practices. In our project, the graph module was constructed using aggregated statistics and optionally with DP training. This makes the approach more suitable for real-world integration into platforms with strict privacy requirements. So, we combine the effectiveness of propagation models with the realities of legal application.

Shen et al. showed that contrastive learning combined with optimal transport (OT) effectively aligns the distributions of different modalities (in particular, text and visual features) [36]. It also improves the quality of cross-modal embeddings. We also used contrastive pre-training and, if necessary, an OT regularizer. However, our contribution is to demonstrate the practical effectiveness of this combination in a real three-component pipeline and in the context of a problem requiring legal guarantees. We empirically show an improvement in portability compared to conventional supervised fusion, which is consistent with the findings of Shen et al., but extends their results to scenarios with additional propagation channels and explainability needs. The research hypothesis, which was that combining multimodal models with a graph module would increase the efficiency and transparency of disinformation detection systems, was confirmed.

The obtained results have practical significance for the creation of effective mechanisms for legal regulation of the digital information space, in particular in the field of countering disinformation. The proposed multimodal model with contrastive prior learning, explainability module and human-in-the-loop procedure can be integrated into content monitoring systems of social networks and news platforms. It is able to perform the function of a tool for preliminary assessment of the reliability of information. Its use enables:

- Increasing the transparency of the automated moderation process, as explainability mechanisms ensure the reproducibility and legal justification of the decisions made;

- Reducing the risks of restricting freedom of speech, as the system does not make a final decision on the removal of content, but generates an explained analytical conclusion that is subject to verification by a human moderator or an independent supervisory body;

- Contributing to the standardization of approaches to content verification, ensuring the unity of technical and legal criteria for evaluating information materials;

Therefore, the results of the study can be used as a theoretical and methodological background for developing regulatory policies in the field of information security, and as a technological component in creating platforms for automated auditing of content moderation algorithms.

5.1. Research limitations

Despite the high performance indicators, the study has a number of limitations. First, the datasets used do not fully reflect the real information environment, where disinformation is more heterogeneous and context-dependent. Second, the multimodal approach required a trade-off between quality and computational complexity, which may limit its application in scalable systems. Third, the study focused on technical and legal aspects, but did not cover the socio-psychological impact of algorithmic moderation on users. Finally, the long-term generalizability of the models remains open, as disinformation is constantly changing forms and distribution strategies.

5.2. Recommendations

Further research may follow several directions. It is worth creating new, more representative datasets that take into account cultural, linguistic, and thematic differences in disinformation. The development of energy-efficient and optimized models suitable for implementation in large social platforms in real time is promising. It is appropriate to deepen the integration of explainability modules and appeal mechanisms to guarantee compliance with the legal principles of transparency and proportionality. Finally, further research should include interdisciplinary collaboration to create more comprehensive systems for combating disinformation that not only effectively detect fakes, but also respect freedom of speech.

6. CONCLUSIONS

The relevance of the problem under study is determined by the rapid spread of disinformation in the digital environment and the urgent need for effective detection tools that maintain the balance between societal security and fundamental human rights. In this study, a comprehensive three-component model was developed and

experimentally validated. It integrates textual data (BERT-base), visual features (ResNet-50), and graph-based content dissemination structures (GNN). Empirical results confirmed the working hypothesis: the integration of multimodal features utilizing progressive fusion and contrastive pre-training achieved high performance metrics (Accuracy ≈ 0.938, F1-score ≈ 0.925, AUC ≈ 0.965), significantly outperforming unimodal and bimodal counterparts.

An important contribution of this research is the implementation of privacy-protection methods based on aggregated statistics and differentially private learning, as well as the modeling of human-in-the-loop scenarios. The significance of these results is that, from a legal perspective, they demonstrate the feasibility of implementing automated moderation without the risks of censorship or mass challenges to platform actions –a solution that fully aligns with the requirements of international instruments, such as the EU Digital Services Act.

The proposed model can serve as a practical tool for moderators and regulatory bodies of digital platforms, striking a clear balance between maintaining a safe environment and protecting freedom of speech. Further research will focus on developing larger, linguistically and culturally annotated multilingual datasets. Another promising direction involves the computational optimization and compression of the proposed architecture through techniques such as quantization or distillation to enable its energy-efficient, real-time deployment on large-scale social platforms.

Further research provides for a deep interdisciplinary collaboration to more precisely integrate explainability modules into the legal and regulatory practices of digital platforms, ensuring strict adherence to the principles of proportionality and personal data protection.

REFERENCES

[1]. U. Nikonenko, T. Shtets, A. Kalinin, I. Dorosh, and L. Sokolik, “Assessing the policy of attracting investments in the main sectors of the economy in the context of introducing aspects of industry 4.0”, *International Journal of Sustainable Development and Planning*, Vol. 17, No. 2, 2022, pp. 497–505, doi: 10.18280/ijsdp.170214

[2]. O. Borodina, H. Kryshtal, M. Hakova, T. Neboha, P. Olczak, and V. Koval, “A conceptual analytical model for the decentralized energy-efficiency management of the national economy”, *Polityka Energetyczna*, Vol. 25, No. 1, 2022, pp. 5–22, doi: 10.33223/epj/147017

[3]. M. Kryshtanovych, L. Akimova, V. Shamrayeva, M. Karpa, and O. Akimov, “Problems of European integration in the construction of EU security policy in the context of counter-terrorism”, *International Journal of Safety and Security Engineering*, Vol. 12, No. 4, 2022, pp. 501-506. doi:10.18280/ijsse.120411 25

[4]. O. Pollicino (Ed.), *Freedom of Speech and the Regulation of Fake News*. Cambridge: Intersentia, 2023.

[5]. L. Hasimi, and A. Poniszewska-Maranda, “Detection of disinformation and content filtering using machine learning: Implications to human rights and freedom of speech”, *ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2024: the 46th European Conference on Information Retrieval)*. Glasgow, UK: ACM, 2024, pp. 68-77

[6]. A. Turillazzi, M. Taddeo, L. Floridi, and F. Casolari, “The digital services act: An analysis of its ethical, legal, and social implications”, *Law, Innovation and Technology*, Vol. 15, No. 1, 2023, pp. 83-106, doi: 10.1080/17579961.2023.2184136?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[7]. I. Buri, and J. van Hoboken, *The Digital Services Act (DSA) Proposal: A Critical Overview*. Institute for Information Law (IViR), University of Amsterdam: Digital Services Act (DSA) Observatory, 2021. Available from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf

[8]. S. Chesterman, “Through a glass, darkly: Artificial intelligence and the problem of opacity”, *The American Journal of Comparative Law*, Vol. 69, No. 2, 2021, pp. 271-294, doi: 10.1093/ajcl/avab012

[9]. F. C. C. Santos, “Artificial intelligence in automated detection of disinformation: A thematic analysis”, *Journalism and Media*, Vol. 4, No. 2, 2023, pp. 679-687, doi: 10.3390/journalmedia4020043

[10]. V. L. Rubin, *Misinformation and Disinformation: Detecting Fakes with the Eye and AI.* Cham: Springer Nature, 2022.

[11]. M. Berrondo-Otermin, and A. Sarasa-Cabezuelo, "Application of artificial intelligence techniques to detect fake news: A review", *Electronics*, Vol. 12, No. 24, 2023, 5041, doi: 10.3390/electronics12245041

[12]. M. A. Al-Asadi, and S. Tasdemir, "Using artificial intelligence against the phenomenon of fake news: A systematic literature review", *Combating Fake News with Computational Intelligence Techniques*. Cham: Springer, 2021, pp. 39-54.

[13]. H. R. Saeidnia, E. Hosseini, B. Lund, M. A. Tehrani, S. Zaker, and S. Molaei, "Artificial intelligence in the battle against disinformation and misinformation: A systematic review of challenges and approaches", *Knowledge and Information Systems,* Vol. 67, No. 4, 2025, pp. 3139-3158, doi: 10.1007/s10115-024-02337-7

[14]. B. Cartwright, R. Frank, G. Weir, and K. Padda, "Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks", *Neural Computing and Applications,* Vol. 34, No. 18, 2022, pp. 15141-15163, doi: 10.1007/s00521-022-07296-0

[15]. O. Pollicino, "General report: Freedom of speech and the regulation of fake news", *Freedom of Speech and the Regulation of Fake News*. Cambridge: Intersentia. 2023, pp. 1-38

[16]. L. Gielow Jacobs, "Freedom of speech and regulation of fake news", *The American Journal of Comparative Law*, Vol. 70, No. Supplement_1, 2022, pp. i278-i311, doi: 10.1093/ajcl/avac010

[17]. D. Vese, "Governing fake news: the regulation of social media and the right to freedom of expression in the era of emergency", *European Journal of Risk Regulation*, Vol. 13, No. 3, 2022, pp. 477-513. Doi: 10.1017/err.2021.48

[18]. N. Bontridder, and Y. Poullet, "The role of artificial intelligence in disinformation", *Data & Policy*, Vol. 3, 2021, e32, doi: 10.1017/dap.2021.20

[19]. P. R. Banchio, "Legal framework to combat disinformation and hate apeech on digital platforms", *SSRN,* 2024, 4879162, doi: 10.2139/ssrn.4879162

[20]. G. De Gregorio, and P. Dunn, "Artificial intelligence and freedom of expression", Artificial Intelligence and Human Rights. Oxford: Oxford University Press, 2023, pp. 76-90.

[21]. B. Azgin, and S. Kiralp, "Surveillance, disinformation, and legislative measures in the 21st century: AI, social media, and the future of democracies", *Social Sciences*, Vol. 13, No. 10, 2024, 510, doi: 10.3390/socsci13100510

[22]. F. Fusco, "Artificial intelligence and security: Some reflections concerning the freedom of expression, information and democracy", *International Journal of Electronic Security and Digital Forensics*, Vol. 17, No. 4, 2025, pp. 500-509.

[23]. C. Gilbert, and M. A. Gilbert, "The role of artificial intelligence (AI) in combatting deepfakes and digital misinformation", *International Research Journal of Advanced Engineering and Science*, Vol. 9, No. 4, 2024, pp. 170-181.

[24]. A. Montoro Montarroso, J. Cantón-Correa, and J. Gómez Romero, "Fighting disinformation with artificial intelligence: Fundamentals, advances and challenges", *Profesional de la Información*, Vol. 32, No. 3, 2023, e320322, doi: 10.3145/epi.2023.may.22

[25]. P. Akhtar, A. M. Ghouri, H. U. R. Khan, M. Amin ul Haq, U. Awan, N. Zahoor, Z. Khan, and A. Ashraf, "Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions", *Annals of Operations Research*, Vol. 327, No. 2, 2023, pp. 633-657, doi: 10.1007/s10479-022-05015-5

[26]. A. Iqbal, K. Shahzad, S. A. Khan, and M. S. Chaudhry, "The relationship of artificial intelligence (AI) with fake news detection (FND): A systematic literature review", *Global Knowledge, Memory and Communication*, Vol. 74, No. 5-6, 2025, pp. 1617-1637, doi: 10.1108/GKMC-07-2023-0264?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[27]. R. K. Kaliyar, A. Goswami P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach", *Multimedia Tools and Applications*, Vol. 80, No. 8, 2021, pp. 11765-11788, doi:10.1007/s11042-020-10183-2

[28]. S. Alonso-Bartolome, and I. Segura-Bedmar, "Multimodal fake news detection", *Expert*

Systems with Applications, 2021, arXiv preprint arXiv:2112.04831

[29]. J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks", *Information Processing & Management,* Vol. 60, No. 1, 2023, 103120, doi: 10.1016/j.ipm.2022.103120

[30]. H. Wang, J. Guo, S. Liu, P. Chen, and X. Li, "SEPM: Multiscale semantic enhancement-progressive multimodal fusion network for fake news detection", *Expert Systems with Applications,* 2025, 127741.

[31]. A. M. Luvembe, W. Li, S. Li, G. Xu, X. Wu, and F. Liu, "An adaptive auto fusion with hierarchical attention for multimodal fake news detection", *Expert Systems with Applications,* 2025, 127930, doi: 10.1016/j.eswa.2025.127930

[32]. Z. Guo, K. Yu, A. Jolfaei, G. Li, F. Ding, and A. Beheshti, "Mixed graph neural network-based fake news detection for sustainable vehicular social networks", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 12, 2022, pp. 15486-15498.

[33]. K. Li, B. Guo, J. Liu, J. Wang, H. Ren, F. Yi, and Z. Yu, "Dynamic probabilistic graphical model for progressive fake news detection on social media platform", *ACM Transactions on Intelligent Systems and Technology (TIST),* Vol. 13, No. 5, 2022, pp. 1-24, doi: 10.1145/3523060

[34]. S. Gong, R. O. Sinnott, J. Qi, and C. Paris, "Fake news detection through graph-based neural networks: A survey", 2023, arXiv preprint arXiv:2307.12639.

[35]. B. Lakzaei, M. Haghir Chehreghani, and A. Bagheri, "Disinformation detection using graph neural networks: A survey", *Artificial Intelligence Review*, Vol. 57, No. 3, 2024, 52, doi: 10.1007/s10462-024-10702-9

[36]. X. Shen, M. Huang, Z. Hu, S. Cai, and T. Zhou, "Multimodal fake news detection with contrastive learning and optimal transport", *Frontiers in Computer Science*, Vol. 6, 2024, 1473457, doi: 10.3389/fcomp.2024.1473457