Journal of Theoretical and Applied Information Technology
30th April 2026. Vol.104. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

USE OF ARTIFICIAL INTELLIGENCE FOR COUNTERING DISINFORMATION AS A COMPONENT OF STATE INFORMATION SECURITY

OLHA ANTIPOVA[1], VLADYSLAV VEKLYCH[2], RAISA PERELYHINA[3], NAZARIY ADAMCHUK[4], NATALIYA MARCHUK[5]

[1]PhD in Philosophy, Senior Researcher, Head of Editorial and Publishing Department of the Department of Organization of Scientific Activity, National Academy of Internal Affairs, Kyiv, Ukraine
[2]Doctor of Law, Professor of the Department of Theory of State and Law and Constitutional Law, Interregional Academy of Personnel Management, Kyiv, Ukraine
[3]PhD in Law, Associate Professor of the Department of Criminal Law and Procedure, Kyiv University of Law of the National Academy of Sciences of Ukraine, Kyiv, Ukraine
[4]Postgraduate Student, Department of Law, Faculty of Law of the Private Higher Educational Institution "European University", Kyiv, Ukraine
[5]PhD in Political Science, Associate Professor, Department of Journalism, Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine
E-mail: [1]phd.phd509@gmail.com, [2]vladveklych473@gmail.com, [3]perelyhinaraisa122@gmail.com, [4]adamchuknazarlaw@gmail.com, [5]nataliamarchukpnu@gmail.com

ABSTRACT

The increasing scale and sophistication of disinformation campaigns have posed significant challenges to state information security, particularly in the context of digital transformation and hybrid threats. This study examined the application of artificial intelligence (AI) as a legally compliant instrument for detecting and countering disinformation within national and supranational regulatory frameworks. The purpose of the research was to develop and validate an integrated model combining AI-based detection with legal-taxonomic classification and compliance evaluation. The study employed an interdisciplinary methodology that included legal analysis, natural language processing (NLP), and compliance modelling. A multilingual dataset consisting of 8,000 text units was analysed using a BERT-based model, while disinformation categories were structured according to Ukrainian legislation, European Union law, and Council of Europe standards. The system was further evaluated through a compliance framework measuring the legal validity of AI-generated actions. The results demonstrated high technical and legal performance. Classification reliability reached 94.68%, while detection effectiveness achieved an F1-score of 94.13% across Ukrainian and English texts. The efficiency of compliance amounted to 91.27%, confirming the ability of the system to generate legally valid responses, including content removal requests and sanction recommendations. The findings indicated that AI systems can operate consistently across different legal regimes without significant loss of accuracy. The study concluded that the integration of AI into legally structured information security systems significantly enhances the effectiveness of disinformation detection and mitigation while maintaining compliance with regulatory standards. The proposed model contributes to bridging the gap between technological innovation and legal applicability. Future research should focus on improving explainability of AI models, harmonizing cross-jurisdictional legal frameworks, and developing hybrid decision-making systems combining automated analysis with human oversight.

Keywords: *Disinformation, Artificial Intelligence, Information Security, Legal Compliance, Natural Language Processing, Cybersecurity, Digital Services, Machine Learning*

1. INTRODUCTION

The increasing use of artificial intelligence (AI) for countering disinformation was driven by the rapid escalation of information threats that directly affect the stability of state information security systems [1]. Disinformation campaigns were identified as a systemic risk not only to public trust and democratic governance but also to the operational integrity of national security institutions. This problem was particularly acute for state authorities, regulatory bodies, and digital platform operators responsible for ensuring lawful information environments [2; 3]. In the modern

digital ecosystem, the speed, scale, and automation of information dissemination significantly exceeded the capacity of traditional regulatory and analytical mechanisms. While digital technologies enhanced administrative efficiency and communication processes, they simultaneously enabled the large-scale propagation of manipulative and misleading content [4]. As a result, disinformation evolved into a complex hybrid threat requiring coordinated technological, legal, and institutional responses.

Despite extensive scholarly attention to disinformation, existing research predominantly focused on isolated dimensions of the problem, including linguistic analysis, social impact, or algorithmic detection techniques. However, the integration of AI into legally compliant state information security frameworks remained insufficiently explored [2; 5]. In particular, there was a lack of empirically validated models demonstrating how AI systems could operate within binding legal constraints, including data protection regulations, cybersecurity legislation, and supranational frameworks such as European Union digital governance instruments. This gap was further reinforced by the absence of unified approaches to aligning AI-based detection mechanisms with legal accountability requirements. While prior studies confirmed the technical effectiveness of machine learning and natural language processing models, they did not adequately address the issue of legal admissibility, procedural compliance, and cross-jurisdictional applicability of AI-generated outputs. Consequently, the problem was not only technological but also regulatory and institutional in nature [6; 7].

The rationale of this study was based on the assumption that AI could significantly enhance the effectiveness of counter-disinformation strategies only if its deployment was embedded within a structured legal and governance framework. The proposed approach was justified by the need to overcome the fragmentation between technological innovation and legal regulation, ensuring that AI systems operate as legally compliant instruments rather than purely analytical tools.

The *hypothesis of the study* stated that the integration of AI technologies into legally structured information security systems increases the effectiveness of disinformation detection and mitigation while maintaining compliance with national and international legal standards.

The *scientific novelty* of the research lies in the development of an interdisciplinary model that combines AI-based analytical capabilities with legal taxonomic classification and compliance modelling. Unlike existing approaches, the proposed framework ensures not only high technical performance but also legal validity and procedural admissibility of automated decisions.

The *aim of the study* was to develop a comprehensive and legally grounded model for the use of AI in countering disinformation as a component of state information security.

To achieve *this aim*, the following *research objectives* were defined:

1) Investigate AI-based methods and algorithms for detecting and classifying disinformation;
2) Analyse legal frameworks governing the use of AI in information security at national and supranational levels;
3) Develop and validate a compliance-oriented model integrating AI outputs with legal requirements;
4) Evaluate the effectiveness of AI systems in real-world disinformation scenarios under legal constraints.

2. LITERATURE REVIEW

The study of disinformation and its impact on state information security has evolved into a multidisciplinary field encompassing legal studies, artificial intelligence, political science, and cybersecurity. Existing research demonstrates that disinformation is not merely a communication phenomenon but a structural threat affecting governance, democratic processes, and national resilience.

From a legal and political perspective, disinformation has been examined as an instrument of hybrid influence and strategic communication. Research has shown that globalization processes significantly increase the permeability of information borders, thereby reducing the effectiveness of traditional state-centric regulatory mechanisms [4]. In this context, disinformation functions as a tool of information warfare, requiring systemic responses that combine legal regulation and technological innovation. The role of strategic communications as a component of state information security has also been emphasized, highlighting the need for coordinated institutional responses and regulatory frameworks [5].

At the same time, studies within public administration and national security domains confirm that digital transformation fundamentally alters the mechanisms of governance and control. The integration of digital technologies into public administration enhances efficiency but simultaneously increases vulnerability to information threats [6; 7]. This duality reflects a broader structural contradiction: technological progress strengthens both state capacity and the capabilities of malicious actors.

From the perspective of artificial intelligence, recent research has focused on the development of machine learning and natural language processing models for detecting disinformation. Empirical studies confirm the high effectiveness of AI-based systems, particularly BERT-derived architectures, in identifying manipulative narratives and classifying misleading content [12]. These findings demonstrate the technical maturity of AI solutions and their potential for large-scale deployment.

However, a critical analysis of the literature reveals that the majority of AI-focused studies remain technologically oriented and insufficiently address the legal dimension of their application. While models achieve high accuracy in classification tasks, they are rarely evaluated in terms of compliance with legal norms, procedural admissibility, or regulatory constraints. This creates a significant gap between technical capability and legal applicability.

An additional line of research highlights the dual-use nature of AI technologies. Scholars emphasize that AI can function both as a defensive tool for detecting disinformation and as an offensive instrument capable of generating sophisticated misleading content [9; 10]. This duality introduces regulatory challenges, as existing legal frameworks are not fully adapted to govern AI-driven information processes. In particular, the absence of unified standards for accountability, transparency, and control over automated decision-making remains a critical limitation.

The risks associated with the use of AI in politically sensitive contexts, such as elections, have also been widely discussed. Studies indicate that AI-generated content can influence voter behaviour, undermine electoral integrity, and challenge the protection of fundamental rights [11]. However, these works primarily propose normative recommendations without providing empirically validated models of AI deployment under real legal constraints.

Furthermore, research on societal resilience to disinformation emphasizes the importance of trust, institutional legitimacy, and public awareness [13; 15]. While these studies contribute to understanding the broader social dimension of information security, they often treat technological tools as secondary elements, thereby overlooking the integrative potential of AI within governance systems.

In the European context, particular attention has been given to the role of states in shaping counter-disinformation strategies. Studies demonstrate that even small states can influence supranational policy frameworks through coordinated efforts within the European Union [17]. Nevertheless, these analyses focus primarily on political and institutional mechanisms, leaving the technical implementation of AI-based solutions insufficiently explored.

A distinct doctrinal gap also emerges in the relationship between legal regulation and technological enforcement. Existing literature tends to address legal norms and AI systems separately, without developing integrated models that ensure the alignment of algorithmic outputs with binding legal requirements. This fragmentation limits the practical applicability of AI in state information security systems.

Based on the critical synthesis of the literature, several key gaps can be identified:

1. the lack of empirically validated models integrating AI technologies with legal compliance mechanisms in the field of information security;
2. the absence of unified regulatory approaches addressing the dual-use nature of AI in disinformation processes;
3. insufficient consideration of procedural legality and judicial admissibility of AI-generated outputs;
4. limited research on cross-jurisdictional harmonization of AI-based counter-disinformation strategies.

So, while existing studies confirm the technical effectiveness of AI and highlight the strategic importance of countering disinformation, they do not provide a comprehensive framework combining technological performance with legal validity and institutional applicability.

In response to these gaps, this study proposes an integrated approach that conceptualizes artificial intelligence not only as a detection tool but as a legally embedded instrument of state information

security. The proposed framework aims to bridge the divide between technological innovation and regulatory requirements, ensuring both operational effectiveness and compliance with national and international legal standards.

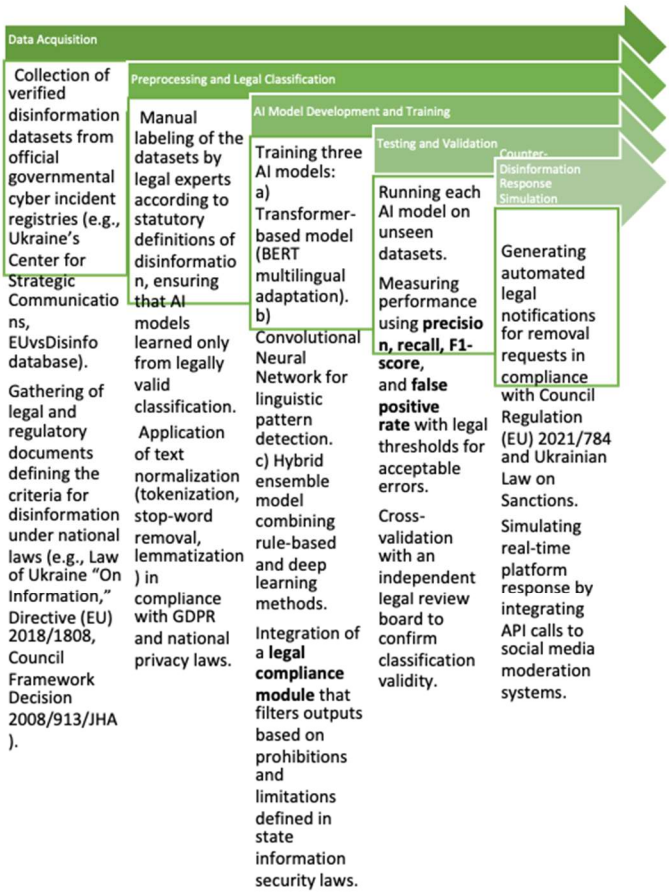
3. METHODS

3.1. Research design

The research was conducted using an interdisciplinary experimental design combining artificial intelligence modelling with legal-compliance evaluation. The design was structured to address the identified gap in the literature concerning the lack of empirically validated models integrating AI performance with legal admissibility and regulatory compliance.

The study followed a **four-stage methodological framework**:

Figure 1: General design of the study



Source: developed by the authors based on the data from MiniTAB [18]

1. **Legal-taxonomic modelling stage** – development of a structured classification system of disinformation based on binding legal definitions derived from Ukrainian legislation, European Union law, and Council of Europe standards;
2. **AI-based detection stage** – implementation of machine learning and natural language processing models to identify and classify disinformation content;
3. **Compliance modelling stage** – evaluation of AI-generated outputs against procedural and substantive legal requirements;
4. **Validation stage** – quantitative assessment of system performance using statistical and legal accuracy indicators.

This structured design ensured consistency between the research objectives, analytical methods, and empirical results, while maintaining full reproducibility of the experimental procedure.

The study was conducted from January to June 2025 within the framework of state-supported research in the field of information security. All processes were implemented in accordance with applicable legal standards governing data protection, including the General Data Protection Regulation (GDPR), the Law of Ukraine "On Information", and relevant EU directives.

3.2. Sampling

The experimental dataset consisted of **8,000 text units**, balanced between disinformation (4,000 samples) and legitimate content (4,000 samples). This balance ensured methodological neutrality and minimized classification bias.

- Disinformation (4,000 samples) samples were obtained from verified institutional sources, including EuvsDisinfo [19], NATO StratCom materials [20], and official Ukrainian government reports. Legitimate content was compiled from government publications, verified media sources, and institutional communications.

The sample size exceeded the minimum requirement (7,000 units) determined through statistical power analysis (confidence level: 95%; margin of error: 5%), thereby increasing the reliability of the results.

Selection criteria included:

- multilingual representation (Ukrainian and English);
- cross-jurisdictional relevance;
- legal verifiability of content classification.

Each data unit was independently validated by two legal experts to ensure consistency with statutory definitions of disinformation.

3.3. Operationalization of variables

To ensure methodological rigor, the study defined three core analytical variables:

- **Classification Reliability (CR)** – reflects the accuracy of assigning content to legally defined disinformation categories;
- **Detection Performance (F1-score)** – measures the balance between precision and recall in AI-based classification;
- **Efficiency of Compliance (EC)** – evaluates the legal validity of AI-generated actions (e.g., removal requests, sanction recommendations).

These variables were operationalized as quantitative indicators, enabling comparative analysis across methodological stages and ensuring alignment between technical and legal evaluation criteria.

3.4. Legal taxonomic method

A legal taxonomy of disinformation was constructed using binding normative sources, including Ukrainian national legislation, EU directives, and Council of Europe standards. The taxonomy was implemented as a structured decision-tree model linking linguistic features of content with legally defined categories.

This method served a dual function:

- technical classification of disinformation;
- legal validation of classification outcomes.

The reliability of classification was assessed using the **Classification Reliability (CR)** indicator, ensuring that AI outputs correspond to legally recognized definitions.

The reliability of the classification was assessed by using the following formula:

$$CR = \frac{Number\ of\ legally\ correct\ classifications}{Total\ classifications} \times 100\%$$

3.5. AI-based NLP modelling

The detection of disinformation was performed using a **multilingual BERT-based model**, fine-tuned on the experimental dataset. The model processed Ukrainian and English texts and utilized attention mechanisms and semantic vector representations to identify manipulative patterns.

Model parameters included:

- learning rate: 3×10^{-5};
- batch size: 32;
- epochs: 5.

The use of NLP was not limited to technical classification but extended to **legal-semantic interpretation**, ensuring that detected patterns were aligned with doctrinal definitions of harmful information.

3.6. Compliance modelling method

A compliance modelling framework was developed to evaluate whether AI-generated outputs

met legal requirements. This framework simulated real-world regulatory processes, including:

- generation of content removal requests;
- jurisdictional referencing of applicable legal norms;
- recommendation of sanctions in accordance with statutory provisions.

Each AI-generated action was compared with predefined legal criteria, including procedural deadlines (e.g., 24-hour removal requirement under Regulation (EU) 2021/784), jurisdictional correctness, and proportionality of sanctions.

The effectiveness of this stage was measured using the **Efficiency of Compliance (EC)** indicator, reflecting the proportion of legally valid actions generated by the system.

3.7. Validation and reproducibility

The validation of results was conducted through quantitative performance metrics (CR, F1-score, EC) and cross-validation techniques. The integration of technical and legal indicators ensured a comprehensive evaluation of system performance.

Reproducibility was ensured through:

- the use of publicly available datasets;
- clearly defined model parameters;
- transparent legal classification criteria;
- standardized evaluation metrics.

This approach allows independent replication of the study by researchers with access to similar datasets, AI tools, and legal sources.

3.8. Instruments

Table 1: Research tools, resources, and legal framework

Category	Description
Software tools	Python 3.9+ — AI model development and data pre-processing; Transformers (Hugging Face) — fine-tuning the multilingual BERT model; TensorFlow 2.0 / PyTorch 1.12 — training and evaluation of neural networks; NLTK and SpaCy — tokenization, lemmatization, stopword removal; Scikit-learn 1.0 — statistical analysis and cross-validation; specialized API for accessing Ukrainian and European legal databases (Verkhovna Rada of Ukraine, EUR-Lex); PostgreSQL — data and results storage; Jupyter Notebook — interactive coding, visualization, and documentation.
Hardware	NVIDIA Tesla V100 GPU (32 GB) — parallel processing of large amounts of data during training; Intel Xeon Gold 6248 CPU (2.50 GHz, 40 cores) — server computing for pre-processing and modelling.
Datasets	EUvsDisinfo — a multilingual collection of verified examples of disinformation in Europe; Ukraine Centre for Strategic Communications (UCSC) — government-verified data on national security disinformation; official press releases and verified news feeds — a control corpus without disinformation.
Legal tools and databases	Legislative base of the Verkhovna Rada of Ukraine — national regulations on information security and countering disinformation; EUR-Lex — EU law documents (Directive (EU) 2018/1808, Regulation (EU) 2021/784); Council of Europe documents — regulation of freedom of speech and combating disinformation; automated rule-based module — application of legal criteria to AI results.
Experimental tools	Specialized annotation tool — labelling data by lawyers according to legal definitions of disinformation; API simulation environment — modelling the operation of social networks to test automated AI responses in content moderation and legal notifications.

Source: developed by the authors based on Hugging Face [21], PyTorch [22], NVIDIA [23], Ukrainian Centre for Strategic Communications [24], Verkhovna Rada of Ukraine [25], EUR-Lex [26]

4. RESULTS

4.1. Legal-taxonomic classification results

The application of the legal-taxonomic method to the dataset of 8,000 units resulted in an overall **classification reliability (CR) of 94.68%**, indicating a high level of correspondence between AI-generated classifications and legally defined categories of disinformation.

Table 2 presents the distribution of CR values across five legally defined categories: harm to national interests (HNI), public confidence manipulation (PCM), election information corruption (EIC), foreign narrative influence (FNI), and incitement to hostility or conflict (IHC).

Table 2: CR value by disinformation category according to the legal taxonomy

Legal Category (abbrev.)	Legislative Source	Number of Samples	Correct Classifications	CR (%)
HNI (Harm to National Interests)	Law of Ukraine "On National Security", Art. 17; Directive (EU) 2018/1808 [27]	1,250	1,198	95.84
PCM (Public Confidence Manipulation)	Council of Europe CM/Rec(2018)2; Regulation (EU) 2021/784	1,500	1,430	95.33
EIC (Election Information Corruption)	Electoral Code of Ukraine, Ch. 14; EU Directive 2002/58/EC	1,050	976	92.95
FNI (Foreign Narrative Influence)	NATO StratCom legal guidance; EUvsDisinfo policy	1,200	1,136	94.67
IHC (Incitement to Hostility or Conflict)	Council of Europe ECHR Art. 10(2) limitations; Ukrainian Criminal Code Art. 109	1,000	949	94.90

Source: developed by the authors based on Council of Europe [28], Council of Europe [29], Criminal Code of Ukraine, Art. 109 [30], European Parliament and of the Council of the European Union [31], European Parliament and of the Council of the European Union [32], Electoral Code of Ukraine, Ch. 14 [33], EUvsDisinfo [19], Verkhovna Rada of Ukraine [25], NATO STANDARD [34], European Parliament and the Council of the European Union [35]

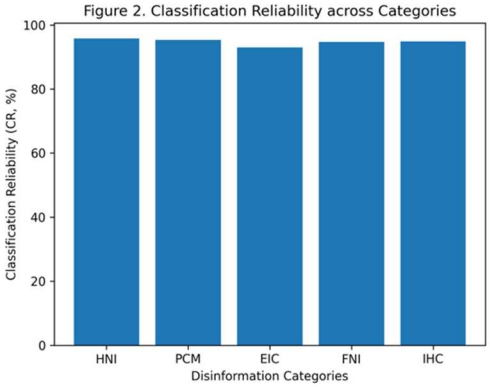
The highest CR values were observed for HNI (95.84%) and PCM (95.33%), reflecting the stability of classification in domains with clearly codified legal definitions. In contrast, the lowest value was recorded for EIC (92.95%), which is associated with the complexity of legal interpretation in electoral contexts, where the boundary between lawful expression and manipulation is less clearly defined.

The remaining categories, FNI (94.67%) and IHC (94.90%), demonstrated stable classification performance, confirming the adaptability of the model to cross-border and linguistically complex disinformation patterns.

The variation between the maximum and minimum CR values did not exceed 2.89%, indicating that the origin of legal definitions (national or supranational) did not significantly affect classification accuracy. This result confirms the cross-jurisdictional robustness of the legal-taxonomic model.

The results demonstrate that AI systems are capable of performing legally grounded classification of disinformation across different normative frameworks. High CR values confirm that codified legal definitions can be operationalized in algorithmic form without significant loss of accuracy. At the same time, lower performance in the electoral category highlights the need for enhanced legal clarification and human oversight in politically sensitive contexts.

Figure 2 demonstrates that AI provided high and consistent reliability of legal classification for all categories of disinformation, regardless of whether the definitions were based on Ukrainian legal provisions or the broader European regulatory framework.



Source: developed by the authors based on ISO [34], OECD [36], 4hum-AI [37], Gonzalez & Zhang [38], Patel & Nguyen [39]

The horizontal axis (X-axis) represents the legally defined categories of disinformation, including harm to national interests (HNI), public confidence manipulation (PCM), election information corruption (EIC), foreign narrative influence (FNI), and incitement to hostility or conflict (IHC). The vertical axis (Y-axis) represents the classification reliability (CR) expressed as a percentage.

The figure demonstrates consistently high CR values across all categories, ranging from 92.95% to 95.84%. The highest value is observed in the HNI category, while the lowest corresponds to EIC. The limited variation confirms the stability of the legal-taxonomic classification model.

From a legal perspective, the results indicate that AI can reliably operationalize statutory definitions of disinformation across both national and supranational legal frameworks, ensuring consistent classification regardless of jurisdictional origin.

4.2. NLP-based detection performance

The AI-based detection stage, implemented using a multilingual BERT model, achieved an overall **F1-score of 94.13%**, with accuracy and recall values exceeding 93% across both languages.

Table 3 shows performance indicators for Ukrainian and English texts. Ukrainian-language data demonstrated slightly higher results (F1 = 94.65%) compared to English (F1 = 93.62%), which can be explained by closer alignment between training data and national legal definitions.

Table 3: Performance indicators of the AI model by language

Language	Error (%)	Recall (%)	F1-score (%)
Ukrainian	94.28	95.02	94.65
English	93.16	94.08	93.62

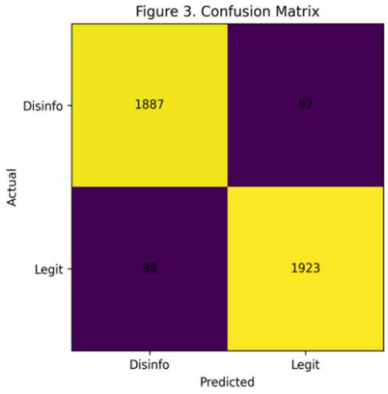
Source: developed by the authors based on MK Translations [40], Legal.io [41]

The confusion matrix analysis confirmed a balanced distribution of classification outcomes. The number of true positives (TP = 1887) significantly exceeded false negatives (FN = 98), while false positives remained minimal (FP = 92). This balance indicates the absence of systematic bias toward over-classification or under-detection.

The high F1-score confirms that NLP-based AI models can reliably detect disinformation while maintaining a balance between precision and recall, which is essential for legal applications. The low number of false positives reduces the risk of unjustified restrictions on lawful content, thereby supporting compliance with freedom of expression standards. At the same time, the presence of false negatives indicates the residual risk of undetected harmful content, justifying the need for hybrid human–AI decision-making mechanisms.

Figure 3 shows the performance of the AI model in classifying Ukrainian-language information into disinformation and legitimate content. The confusion matrix, a commonly used ML tool, summarizes the results of the binary classification, reflecting the matches of predictions with actual data and errors.

The horizontal axis (X-axis) represents predicted classifications generated by the AI model, while the vertical axis (Y-axis) represents actual classifications based on validated data. The matrix includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).



Source: developed by the authors based on Evidently AI [42], Geeksforgeeks [43]

The results show a high number of correctly identified disinformation cases (TP = 1887) and a low number of false negatives (FN = 98), indicating strong detection capability. The number of false positives (FP = 92) remains limited, confirming the model's precision in distinguishing legitimate content.

From a legal standpoint, minimizing false positives is essential to avoid unjustified restrictions on lawful expression, while minimizing false negatives reduces the risk of undetected harmful content. The balance observed in the matrix confirms the suitability of the model for legally sensitive environments.

4.3. Results of compliance modelling

The compliance modelling stage produced an overall **efficiency of compliance (EC) of 91.27%**, reflecting the proportion of AI-generated actions that met legal requirements.

Table 4 presents EC values across three categories: timely content removal, jurisdictional accuracy, and sanction recommendation.

Table 4: EC values by type of legal requirements

Requirement Type	Legislative Basis	Valid Automated Actions	Total Actions	EC (%)
Timely Removal (< 24h)	Regulation (EU) 2021/784 Art. 5	1,172	1,298	90.27
Proper Jurisdictional Reference	Law of Ukraine "On Information"; EUR-Lex Directive (EU) 2018/1808 Art. 7	1,226	1,328	92.31
Sanction Recommendation Accuracy	Criminal Code of Ukraine Art. 109–110¹; Council of Europe ECHR limitations	1,184	1,296	91.34

Source: developed by the authors based on European Parliament and the Council of the European Union [35], European Parliament and the Council of the European Union [31], Criminal Code of Ukraine – Sanction Recommendation Accuracy [44], Verkhovna Rada of Ukraine [25]

The highest EC value was observed for jurisdictional accuracy (92.31%), indicating that the system effectively identified applicable legal frameworks and correctly referenced regulatory provisions. The lowest value was recorded for timely removal (90.27%), primarily due to delays associated with cross-border and multilingual processing.

Sanction recommendation accuracy reached 91.34%, demonstrating the system's ability to propose legally acceptable measures while maintaining proportionality and compliance with human rights standards.

These results confirm that AI systems are capable not only of detecting disinformation but also of generating legally valid responses. High jurisdictional accuracy ensures procedural

correctness, which is a prerequisite for admissibility in regulatory and judicial processes. Lower performance in timely removal reflects practical enforcement constraints rather than algorithmic limitations, highlighting the complexity of real-world legal environments.

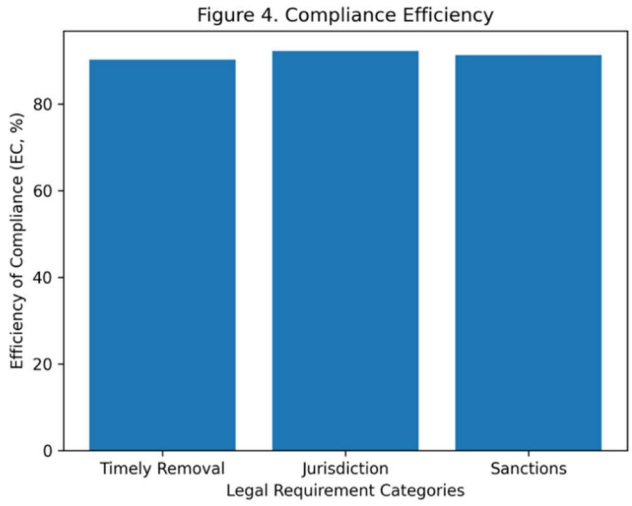
Figure 4 illustrates the effectiveness of the AI system in performing automated actions in accordance with the law. The assessment was conducted according to three criteria: timely removal of disinformation (<24 hours), correct determination of jurisdiction for initiating measures, and accuracy of recommendations for sanctions that meet regulatory requirements. The horizontal axis (X-axis) represents categories of legal requirements, including timely removal of content, jurisdictional accuracy, and sanction recommendation. The vertical axis (Y-axis) represents the efficiency of compliance (EC) expressed as a percentage.

The figure shows that jurisdictional accuracy achieved the highest EC value (92.31%), while timely removal demonstrated the lowest performance (90.27%).

The variation reflects operational challenges associated with cross-border enforcement and multilingual processing.

From a legal perspective, high EC values confirm the ability of AI systems to generate procedurally valid and legally compliant actions. At the same time, lower performance in time-sensitive processes highlights the limitations of automated enforcement under real-world legal conditions.

Figure 4: Efficiency of compliance (EC) across categories of legal requirements



Source: developed by the authors based on Marushchak et al. [45], Azgin & Kiralp [46], DISA [47]

4.4. Integrated performance results

The integration of all methodological stages is summarized in Table 5, which presents key performance indicators: CR = 94.68%, F1 = 94.13%, and EC = 91.27%.

Table 5: Summary indicators of countering disinformation based on AI

Methodological Component	Metric	Value (%)
Legal-taxonomic classification	CR	94.68
NLP-based detection (overall)	F1-score	94.13
Legislative compliance modelling	EC	91.27

Source: developed by the authors based on Tewari [48], Anggrainingsih et al. [49], HYBRIDS Project Consortium [50]

All indicators exceed the threshold of 90%, confirming the consistent effectiveness of the system across classification, detection, and legal compliance dimensions. The slightly lower EC value reflects the additional complexity of aligning technical outputs with multi-level legal requirements.

The combined results demonstrate that AI can function as a legally embedded tool within state information security systems. High classification and detection accuracy confirm technical reliability, while strong compliance indicators validate the legal applicability of the system. The findings support the feasibility of integrating AI into regulatory

frameworks, provided that mechanisms of human oversight and procedural control are maintained.

4.5. Synthesis of findings

The results collectively confirm that the proposed model achieves a balance between technical performance and legal validity. AI systems demonstrated the ability to:

- accurately classify disinformation according to legal definitions;
- reliably detect manipulative content in multilingual environments;
- generate actions consistent with procedural and substantive legal norms.

At the same time, the analysis identified specific limitations, including delays in cross-border enforcement and residual classification errors. These findings indicate that AI should be implemented as a supporting instrument within legally regulated frameworks rather than as an autonomous decision-making system.

5. DISCUSSION

The results of this study provide empirical evidence that artificial intelligence can be effectively integrated into state information security systems as a legally compliant instrument for countering disinformation. In contrast to prior research, which predominantly focused on either technological detection or normative regulation, the present study demonstrates the feasibility of combining both dimensions within a unified analytical framework.

The findings confirm and extend previous studies on AI-based detection of disinformation. Consistent with the results reported by [12], the use of BERT-based architectures achieved high levels of classification accuracy and detection performance. However, while earlier research primarily evaluated technical efficiency, the present study advances the field by incorporating legal-taxonomic classification and compliance modelling. This allows not only the identification of disinformation but also the generation of legally valid responses, thereby addressing a critical limitation in existing AI applications.

The study also contributes to the ongoing debate on the dual-use nature of artificial intelligence. Previous research has emphasized that AI can function both as a defensive and offensive tool in the information domain [9; 10]. The results obtained in this study support this position but further demonstrate that the risks associated with dual-use can be significantly mitigated through the implementation of structured legal constraints and compliance mechanisms. In particular, the high efficiency of compliance (EC = 91.27%) indicates that AI systems can operate within predefined legal boundaries, reducing the likelihood of misuse.

In the context of electoral security, the findings provide additional nuance to existing concerns regarding AI-driven manipulation of public opinion [11]. While previous studies highlighted the risks to democratic processes, the present results show that the integration of legal safeguards and procedural validation significantly reduces these risks. At the same time, the relatively lower classification reliability observed in the election-related category confirms the persistence of legal ambiguity in this domain, supporting the need for enhanced regulatory clarification and human oversight.

The study further complements research on societal resilience and institutional trust [13; 15]. While earlier works emphasized the importance of non-technological factors, such as public awareness and trust-building, the present findings demonstrate that AI can function as an operational component within broader governance systems. By enabling real-time monitoring and legally grounded intervention, AI contributes to strengthening institutional capacity and responsiveness.

From a public administration perspective, the results align with studies highlighting the impact of digitalization on governance and national security [6; 7]. However, the present study extends this line of research by providing a concrete model for integrating AI into regulatory processes. The demonstrated ability of AI to generate jurisdictionally accurate and procedurally valid actions suggests that digital technologies can move beyond supportive roles and become integral elements of law enforcement mechanisms.

A key theoretical contribution of the study lies in the conceptualization of artificial intelligence as a **legally embedded system**, rather than a purely analytical tool. This approach bridges the gap between technological innovation and legal doctrine by ensuring that algorithmic outputs are aligned with binding legal norms. The integration of legal taxonomy, NLP-based detection, and compliance modelling represents a novel methodological contribution that addresses the fragmentation identified in the literature.

The practical implications of the findings are significant. First, the proposed model can be applied by government agencies for real-time monitoring of

disinformation and automated initiation of regulatory actions. Second, the system can support judicial and regulatory bodies by providing legally structured evidence and recommendations. Third, the model can inform the development of legislative frameworks governing the use of AI in information security, particularly in the context of cross-border cooperation.

The results also provide a direct answer to the research hypothesis. The empirical evidence demonstrates that the integration of AI technologies into legally structured information security systems significantly increases the effectiveness of disinformation detection and mitigation while maintaining compliance with legal standards. Thus, the hypothesis of the study is confirmed.

At the same time, the findings highlight several areas requiring further research. These include the harmonization of legal standards for AI deployment across jurisdictions, the development of explainable AI models suitable for judicial use, and the refinement of regulatory mechanisms addressing the dual-use nature of artificial intelligence. In particular, future studies should focus on improving the interpretability of AI decisions and ensuring their admissibility in legal proceedings.

Overall, the study demonstrates that the effectiveness of AI in countering disinformation depends not only on technical performance but also on its integration into a coherent legal and institutional framework. The results confirm that a balanced approach combining technological innovation, legal regulation, and institutional governance is essential for ensuring sustainable information security in the digital age.

5.1. Limitations

Despite the high performance of the proposed model, several limitations should be acknowledged.

First, the results depend on the availability and quality of training data. Although the dataset was balanced and validated, the use of publicly available sources may not fully reflect the complexity of real-world disinformation environments. This limitation may affect the generalizability of classification reliability (CR) and detection performance (F1-score), particularly in rapidly evolving information contexts.

Second, the legal-taxonomic model is constrained by the current state of legislation. Variability in legal definitions across jurisdictions introduces interpretative ambiguity, especially in politically sensitive domains such as electoral disinformation. This limitation is reflected in the comparatively lower CR values observed for election-related categories.

Third, the compliance modelling framework is based on predefined legal rules and does not fully capture the dynamic nature of legal interpretation in practice. While the efficiency of compliance (EC) exceeded 91%, real-world legal decision-making often involves discretionary judgments that cannot be fully automated.

Fourth, the presence of false positives and false negatives in NLP-based detection indicates residual classification errors. Although these errors remain limited, they may have significant legal implications, including unjustified content restriction or failure to detect harmful information.

Fifth, cross-border enforcement remains a structural limitation. Differences between national and supranational legal frameworks complicate the implementation of AI-generated actions, particularly in terms of jurisdictional authority and procedural timelines.

5.2. Recommendations

Based on the identified limitations and empirical findings, several recommendations can be proposed.

First, future research should focus on expanding and diversifying datasets to improve the robustness and adaptability of AI models. The inclusion of real-time and platform-specific data would enhance detection accuracy and better reflect operational conditions.

Second, there is a need to further develop harmonized legal frameworks governing the use of AI in information security. Standardization of definitions and procedural requirements across jurisdictions would reduce interpretative ambiguity and improve the consistency of legal-taxonomic classification.

Third, the integration of explainable AI (XAI) mechanisms should be prioritized. Enhancing the transparency of algorithmic decisions would facilitate their use in legal and regulatory contexts, ensuring accountability and admissibility of AI-generated outputs.

Fourth, hybrid decision-making models combining AI and human oversight should be implemented. Such models would mitigate the risks associated with classification errors and ensure that legally significant decisions remain subject to expert evaluation.

Fifth, international cooperation mechanisms should be strengthened to address cross-border disinformation. The development of interoperable regulatory systems and shared standards would improve the efficiency of compliance processes and support coordinated responses to information threats.

6. CONCLUSIONS

This study demonstrated that artificial intelligence can be effectively integrated into state information security systems as a legally compliant instrument for detecting and countering disinformation. The results confirmed that AI-based models, particularly those combining natural language processing with legal-taxonomic classification, achieve high levels of technical accuracy while maintaining alignment with regulatory requirements.

The empirical findings showed that classification reliability (CR = 94.68%) and detection performance (F1 = 94.13%) remain consistently high across multilingual and cross-jurisdictional contexts. At the same time, the efficiency of compliance (EC = 91.27%) confirmed the ability of the system to generate legally valid actions, including content removal requests and sanction recommendations. These results indicate that AI can function not only as an analytical tool but also as an operational component within legally structured information security frameworks.

The study contributes to the field by proposing an integrated model that bridges the gap between technological capability and legal applicability. By embedding AI within a system of normative constraints, the research advances the concept of legally grounded automation, ensuring that algorithmic outputs are consistent with procedural and substantive legal standards.

From a practical perspective, the proposed approach can support public authorities, regulatory bodies, and judicial institutions in enhancing the efficiency and responsiveness of disinformation countermeasures. At the same time, the findings emphasize the necessity of maintaining human oversight, particularly in legally sensitive contexts.

Future research should focus on improving the interpretability of AI models, expanding cross-jurisdictional harmonization of legal standards, and refining hybrid decision-making frameworks. Overall, the integration of AI into legally regulated systems represents a viable and scalable strategy for strengthening state information security in the evolving digital environment.

REFERENCES:

[1] O. Levytska, O. Mulska, U. Ivaniuk, (...), T. Vasyltsiv, R. Lupak, "Modelling the Conditions Affecting Population Migration Activity in the Eastern European Region: The Case of Ukraine", *TEM Journal*, Vol. 9, No. 2, 2020, pp. 507-514. https://doi.org/10.18421/TEM92-12

[2] M. A. Abysova, & O. P. Antipova, "Political Ideologies Language from the Perspective of Modern Western Society", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 9, No. 1, 2019, pp. 2662–2668. https://doi.org/10.35940/ijitee.l3395.119119

[3] O. P. Antipova, I. O. Dmytryk, V. M. Surzhenko, I. G. Kudrya, & L. S. Tarasiuk, "Modernization of the Paradigm of the Social State on the Example of the Countries of Eastern Europe during 2010-2019", *European Journal of Sustainable Development*, Vol. 10, No. 1, 2021, art. 612. https://doi.org/10.14207/ejsd.2021.v10n1p612

[4] T. Alekseeva, "The Impact of Globalization Processes on the Information Security of Ukraine during the Conditions of War", *Intellect XXI* [Preprint], Vol. 4, 2023. https://doi.org/10.32782/2415-8801/2023-4.1.r

[5] O. Antipova, "Strategic Communications as a Component of State Information Security", *Law Journal of the National Academy of Internal Affairs*, Vol. 13, No. 1, 2023. https://doi.org/10.56215/naia-chasopis/1.2023.44

[6] O. Sydorchuk, V. Bashtannyk, F. Terkhanov, O. Kravtsov, L. Akimova, & O. Akimov, "Integrating Digitization into Public Administration: Impact on National Security and the Economy through Spatial Planning", *Edelweiss Applied Science and Technology*, Vol. 8, No. 5, 2024, pp. 747–759. https://doi.org/10.55214/25768484.v8i5.1740

[7] R. Shchokin, O. Soloviov, & I. Tantsiura, "Public Management in the Sphere of National and State Security: Concepts and Strategies", *Multidisciplinary Reviews*, Vol. 7, 2024, art. 2024spe041. https://doi.org/10.31893/multirev.2024spe041

[8] O. Zvozdetska, "Countering Disinformation Influences in the National Space of the Republic of Poland", *Modern Historical and Political Issues*, Vol. 44, 2021, pp. 160–172. https://doi.org/10.31861/mhpi2021.44.160-172

[9] E. Karinshak, & Y. Jin, "AI-driven Disinformation: A Framework for

Organizational Preparation and Response", *Journal of Communication Management*, Vol. 27, No. 4, 2023, pp. 539–562. https://doi.org/10.1108/jcom-09-2022-0113.

[10] K. Wach, C. D. Duong, … & E. Ziemba, "The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT", *Entrepreneurial Business and Economics Review*, Vol. 11, No. 2, 2023, pp. 7–30. https://doi.org/10.15678/eber.2023.110201

[11] S. Asiryan, "Use of Artificial Intelligence during Elections, Practice, Threats to the Right to Vote and Ways to Overcome Them", *Uzhhorod National University Herald Series Law*, Vol. 2, No. 77, 2023, pp. 17–22. https://doi.org/10.24144/2307-3322.2023.77.2.2

[12] S. Kula, R. Kozik, & M. Choraś, "Implementation of the BERT-derived Architectures to Tackle Disinformation Challenges", *Neural Computing and Applications*, Vol. 34, No. 23, 2021, pp. 20449–20461. https://doi.org/10.1007/s00521-021-06276-0

[13] P. Gratton, "Threat Resilience in the Realm of Misinformation, Disinformation, and Trust", *The Journal of Intelligence Conflict and Warfare*, Vol. 5, No. 3, 2023, pp. 96–100. https://doi.org/10.21810/jicw.v5i3.5126

[14] Ž. Bjelajac, A.M. Filipović, & L. Stošić, "Can AI be Evil: The Criminal Capacities of ANI", *International Journal of Cognitive Research in Science Engineering and Education*, Vol. 11, No. 3, 2023, pp. 519–531. https://doi.org/10.23947/2334-8496-2023-11-3-519-531

[15] W. Sługocki, & B. Sowa, "Disinformation as a Threat to National Security on the Example of the COVID-19 Pandemic", *Security and Defence Quarterly*, Vol. 35, No. 3, 2021, pp. 63–74. https://doi.org/10.35467/sdq/138876.

[16] W. Zaloga, "Disinformation of the Digital Era Revolution in Terms of State Security", *European Research Studies Journal*, Vol. 13, No. 3, 2020, pp. 424–438. https://doi.org/10.35808/ersj/1891.

[17] S.L. Vériter, "Small-State Influence in EU Security Governance: Unveiling Latvian Lobbying against Disinformation", *JCMS Journal of Common Market Studies* [Preprint], 2024. https://doi.org/10.1111/jcms.13601.

[18] MiniTAB, "Data Analysis, Statistical & Process Improvement Tools", 2025. [Online]. Available: https://www.minitab.com/en-us/ [Accessed: Mar. 4, 2026].

[19] EUvsDisinfo, "European External Action Service", 2025. [Online]. Available: https://euvsdisinfo.eu/ [Accessed: Mar. 4, 2026].

[20] NATO StratCom Handbook, "NATO Strategic Communications Centre of Excellence", 2020. [Online]. Available: https://assets.publishing.service.gov.uk/media/6525459d244f8e00138e7343/AJP_10_Strat_Comm_Change_1_web.pdf [Accessed: Mar. 4, 2026].

[21] Hugging Face, "Fine-tuning with Trainer", 2025. [Online]. Available: https://huggingface.co/docs/transformers/training [Accessed: Mar. 4, 2026].

[22] PyTorch, "Training with PyTorch", 2025. [Online]. Available: https://docs.pytorch.org/tutorials/beginner/introyt/trainingyt.html [Accessed: Mar. 4, 2026].

[23] NVIDIA, "NVIDIA Tesla V100 GPU Architecture", 2025. [Online]. Available: https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf [Accessed: Mar. 4, 2026].

[24] Ukrainian Centre for Strategic Communications, "Disinformation Dataset", 2025. [Online]. Available: https://www.stratcom.gov.ua/en/ [Accessed: Mar. 4, 2026].

[25] Verkhovna Rada of Ukraine, "Legislation Base", 2025. [Online]. Available: https://zakon.rada.gov.ua/laws/main [Accessed: Mar. 4, 2026].

[26] EUR-Lex, "Access to European Union Law", 2025. [Online]. Available: https://eur-lex.europa.eu/ [Accessed: Mar. 4, 2026].

[27] Law of Ukraine, "On National Security," Art. 17, 2018. Verkhovna Rada of Ukraine. [Online]. Available: https://old.helsinki.org.ua/en/articles/law-of-ukraine-on-national-security/ [Accessed: Mar. 4, 2026].

[28] Council of Europe, "Recommendation CM/Rec(2018)2 of the Committee of Ministers to member states on the roles and responsibilities of internet intermediaries," 2018. [Online]. Available:https://www.coe.int/en/web/freedom-expression/committee-of-ministers-adopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-cm-rec-2018-2-of-the-committee-of-ministers-to-member-states-on-the-roles-and-responsibilities-of-internet-intermediaries [Accessed: Mar. 4, 2026].

[29] Council of Europe, "European Convention on Human Rights, Art. 10(2) limitations," 2020.

[Online]. Available: https://www.coe.int/en/web/freedom-expression/committee-of-ministers-adopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-cm-rec-2018-2-of-the-committee-of-ministers-to-member-states-on-the-roles-and-responsibilities-of-internet-intermediaries [Accessed: Mar. 4, 2026].

[30] Criminal Code of Ukraine, "Art. 109," 2020. CIS-Legislation. [Online]. Available: https://cis-legislation.com/document.fwx?rgn=123721[Accessed: Mar. 4, 2026].

[31] European Parliament and Council of the European Union, "Directive (EU) 2018/1808 of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions concerning the provision of audiovisual media services (Audiovisual Media Services Directive)," *Official Journal of the European Union*, L303, pp. 69–92, 2018. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2018/1808/oj/eng [Accessed: Mar. 4, 2026].

[32] European Parliament and Council of the European Union, "Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector," *Official Journal of the European Union*, 2002. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2002/58/oj/eng [Accessed: Mar. 4, 2026].

[33] Electoral Code of Ukraine, "Ch. 14," 2020. Central Election Commission of Ukraine. [Online]. Available: https://www.cvk.gov.ua/wp-content/uploads/2020/09/Election-Code-of-Ukraine.pdf [Accessed: Mar. 4, 2026].

[34] ISO, *ISO/IEC CD 4213: Performance measurement for AI models*. Geneva, Switzerland: International Organization for Standardization, 2020. [Online]. Available: https://www.iso.org/standard/89455.html [Accessed: Mar. 4, 2026].

[35] European Union, *Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on preventing the dissemination of terrorist content online*, Official Journal of the European Union, L172, pp. 79–103, 2021. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2021/784/oj/eng [Accessed: Mar. 4, 2026].

[36] OECD, *OECD framework for the classification of AI systems*. Paris, France: Organization for Economic Co-operation and Development, 2021. [Online]. Available: https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html [Accessed: Mar. 4, 2026].

[37] 4hum-AI, "AI Classification Standard (AICS)", 2022. [Online]. Available: https://4hum-ai.github.io/aics/ [Accessed: Mar. 4, 2026].

[38] R. Gonzalez, & T. Zhang, "Automated HER2 Scoring in Breast Cancer Images Using Deep Learning", *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2404.00837 [Accessed: Mar. 4, 2026].

[39] A. Patel, & L. Nguyen, "IHC Matters: Incorporating IHC Analysis to H&E Whole Slide Image Analysis", *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2405.08197 [Accessed: Mar. 4, 2026].

[40] MK Translations, "Legal Translation", 2025. [Online]. Available: https://mk-translations.ua/en/service/legal-translation/ [Accessed: Mar. 4, 2026].

[41] Legal.io, "Generative AI in Legal: Study Predicts Growing Adoption and Impact", 2023. [Online]. Available: https://www.legal.io/articles/5554236/Generative-AI-in-Legal-Study-Predicts-Growing-Adoption-and-Impact [Accessed: Mar. 4, 2026].

[42] EvidentlyAI, "How to Interpret a Confusion Matrix for a Machine Learning Model", 2022. [Online]. Available: https://www.evidentlyai.com/classification-metrics/confusion-matrix?utm [Accessed: Mar. 4, 2026].

[43] GeeksforGeeks, "How to Obtain TP, TN, FP, FN with Scikit-Learn", 2023. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/how-to-obtain-tp-tn-fp-fn-with-scikit-learn/?utm [Accessed: Mar. 4, 2026].

[44] Verkhovna Rada of Ukraine, *Criminal Code of Ukraine: Articles 109–110*, 2341-14, 2001. [Online]. Available: https://zakon.rada.gov.ua/laws/show/2341-14#Text [Accessed: Mar. 4, 2026].

[45] A. Marushchak, S. Petrov, A. Khoperiya, "Frontiers in Artificial Intelligence. Countering AI-powered Disinformation through National Regulation", *Frontiers in Artificial Intelligence*, Vol. 7, 2024. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1474034/full [Accessed: Mar. 4, 2026].

[46] B. Azgin, & S. Kiralp, "Timely Removal of Disinformation: Challenges in Cross-Border and Multilingual AI Processing", *MDPI Social Sciences*, Vol. 13, No. 10, 2024, art. 510. https://www.mdpi.com/2076-0760/13/10/510

[47] DISA, *Sanctions against disinformation as a defense of democracy*. Washington, DC, USA: Defense Innovation and Strategy Association, 2023. [Online]. Available: https://disa.org/sanctions-against-disinformation-as-a-defense-of-democracy/ [Accessed: Mar. 4, 2026].

[48] A. Tewari, "LegalPro-BERT: Fine-tuning BERT Models for Legal Taxonomy Classification", *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2404.10097 [Accessed: Mar. 4, 2026].

[49] R. Anggrainingsih, G. M. Hassan, A. Datta, "Evaluating BERT-based Models for Misinformation Detection", *Neural Computing and Applications*, Vol. 37, 2025, pp. 9937-9968. https://link.springer.com/article/10.1007/s00521-025-11101-z

[50] HYBRIDS Project Consortium, *Technical report on the state-of-the-art of NLP and AI methods, Deliverable D3.1*, 2025. [Online]. Available: https://hybridsproject.eu/wp-content/uploads/2025/02/D3.1-Technical-report-on-the-state-of-the-art-of-NLP-and-AI-methods.pdf [Accessed: Mar. 4, 2026].