

MOLCHANOVA Maryna  
Khmelnitskyi National University  
DUTT Pawan Kumar  
Tallinn Technical University (Estonia)

## ARTIFICIAL INTELLIGENCE APPROACH TO IDENTIFYING PROPAGANDA TECHNIQUES AND OBJECTS, TAKING INTO ACCOUNT ETHICAL AND LEGAL ASPECTS

*The article explores the ethical and legal aspects of applying artificial intelligence (AI) technologies to detect propaganda techniques in textual content. The study presents a multi-level approach to identifying signs of propaganda in textual data, recognizing common rhetorical strategies of influence, and establishing semantic links between the detected techniques and their respective targets. The consistent use of neural network models is justified, as it ensures both classification accuracy and transparency of the obtained results through the application of local interpretability methods. The paper presents experimental results based on a corpus of Ukrainian-language news texts and informational messages from social media platforms. The proposed approach demonstrated alignment between the model's predictions and independent expert assessments, confirming its potential applicability in conditions with limited human oversight.*

*Special attention is given to the compliance of the proposed system with existing regulatory frameworks, including constraints on automated decision-making, the user's right to explanation, and the prevention of discriminatory effects resulting from biased training data. The study addresses risks associated with misclassification, potential impacts on freedom of expression, and the accountability of developers in cases where the system is applied in automated content moderation scenarios.*

*The integration of interpretability tools into neural network analysis is proposed as a core design principle to ensure adherence to ethical AI standards. Based on the obtained findings, the study concludes that the development of such systems requires the simultaneous consideration of technical effectiveness, legal compliance, and social responsibility, which are essential conditions for their safe implementation in the practice of analyzing public communications.*

*Keywords: artificial intelligence, ethical aspects, legal regulation, propaganda detection, natural language processing, neural network models, model interpretability, automated decision-making, information security, content moderation.*

МОЛЧАНОВА Марина  
Хмельницький національний університет  
ДАТТ Паван Кумар  
Талліннський технічний університет (Естонія)

## ПІДХІД ВИКОРИСТАННЯ ЗАСОБІВ ШТУЧНОГО ІНТЕЛЕКТУ ДО ІДЕНТИФІКАЦІЇ ПРИЙОМІВ ТА ОБ'ЄКТІВ ПРОПАГАНДИ З ВРАХУВАННЯМ ЕТИЧНИХ ТА ПРАВОВИХ АСПЕКТІВ

*Стаття присвячена дослідженню етичних та правових аспектів застосування технологій штучного інтелекту (ШІ) для виявлення пропагандистських прийомів у текстовому контенті. У роботі розглядається багаторівневий підхід до виявлення у текстових даних ознак пропаганди, визначення типових риторичних технік впливу та встановлення зв'язків між ідентифікованими прийомами й об'єктами впливу.*

*Обґрунтовано послідовне застосування нейромережових моделей, яке забезпечує як точність класифікації, так і прозорість отриманих результатів за рахунок використання локальної інтерпретації результатів. Наведені результати експериментального дослідження на корпусі україномовних новинних повідомлень та інформаційних повідомлень з соціальних платформ. Запропонований підхід продемонстрував відповідність результатів передбачення оцінкам незалежних експертів, що підтверджує можливість його застосування в умовах обмеженого людського контролю.*

*Особливу увагу приділено відповідності функціонування запропонованого підходу чинному нормативному регулюванню, включаючи вимоги щодо обмеження автоматизованого прийняття рішень, права користувача на пояснення, а також запобігання дискримінаційним ефектам на основі упереджених даних навчання. Розглянуто ризики, пов'язані з хибною класифікацією, потенційним впливом на свободу вираження поглядів, а також відповідальністю розробника у разі використання системи в автоматизованих рішеннях, що стосуються контентної модерації.*

*Запропоновано інтеграції засобів пояснюваності як складової при нейромережевому аналізі, що дозволяє забезпечити дотримання принципів етичного ШІ. На основі отриманих результатів зроблено висновок, що розробка таких систем потребує одночасного врахування технічної ефективності, нормативно-правового супроводу та соціальної відповідальності, що є необхідною умовою їх безпечного впровадження у практику аналізу публічних комунікацій.*

*Ключові слова: штучний інтелект, етичні аспекти, правове регулювання, виявлення пропаганди, обробка природної мови, нейромережові моделі, пояснюваність моделей, автоматизоване прийняття рішень, інформаційна безпека, контентна модерація.*

### Introduction

In today's digital environment, the spread of propaganda via text messages on social networks and news platforms poses a serious threat to information security and societal stability. Thanks to their ability to process large amounts of data and detect hidden patterns, artificial intelligence systems have become an effective tool for automatically detecting propaganda techniques in natural language texts. However, the implementation of such

systems raises a number of ethical and legal issues related to the transparency of algorithms, model bias, respect for human rights and regulatory requirements.

As stated in the requirements of the General Data Protection Regulation (GDPR) [1], an individual is guaranteed the right not to be subject to a decision based solely on automated processing if it significantly affects his or her rights and freedoms (Article 22). The EU AI Act [2] states that systems used to assess or influence public sentiment may be classified as high-risk systems. Such systems must meet the requirements of transparency, explainability, non-discrimination, and provide for the possibility of auditing and appealing automated decisions.

Also, ethical frameworks are defined in documents such as: OECD AI Principles [3], UNESCO Recommendation on the Ethics of Artificial Intelligence [4], Human Centric AI: A Comment on the IEEE's Ethically Aligned Design [5].

In the legislation of Ukraine, there is a lack of a clearly formulated regulatory framework for the use of AI in the field of information security, which creates challenges in adapting European standards to Ukrainian realities. However, at the level of state initiatives, in particular within the framework of the Government Action Plan for 2024 [6], the need to strengthen the capacity to counter information threats has been emphasized.

Thus, modern technical solutions in the detection of propaganda, although they demonstrate high potential, require support by regulatory and ethical mechanisms that ensure a balance between accuracy, transparency and user rights.

The main contribution of the paper is the proposed approach to ensuring transparency and explainability of deep learning model decisions, methods for minimizing algorithmic bias, as well as compliance with legal norms regarding the processing of personal data and automated decision-making. Particular attention is paid to the development of system architecture that combines the effectiveness of propaganda detection with compliance of ethical principles and legal requirements.

Further, the structure of the paper is as follows: the section «Literature review» provides an overview of the current state of the scientific direction of responsible and explained artificial intelligence in terms of solving the problem of detecting propaganda influences; the section «Proposed approach» provides an approach to implementing multi-level processing of text content to detect propaganda techniques and corresponding objects of influence; the section «Results and discussion» presents the results of an experimental study of the effectiveness of the developed approach on Ukrainian-language text corpora, including metrics of classification accuracy, interpretation quality and compliance of conclusions with experts' expectations, and also discusses the feasibility of practical application of the system in conditions of increased ethical requirements; The final section «Conclusions» summarizes the main scientific provisions of the study, outlines the potential of the proposed approach for further research in the field of responsible artificial intelligence and its use in the field of information security.

### Literature review

Much of the current research on propaganda detection in natural language texts is based on the application of deep learning methods and transformative architecture models, such as BERT, RoBERTa, and DeBERTa. In particular, within the framework of the SemEval-2020 Task 11, it was proposed to classify 14 propaganda techniques in news content, which became the basis for many subsequent approaches to the automated detection of manipulative techniques. In [7] and [8], it is noted that deep models demonstrate high accuracy, but are limited in the explainability of their decisions. The authors of [9] investigate the vulnerability of pre-trained language models, such as BERT, to attacks using deliberate text modification aimed at manipulating the results of propaganda detection. The main attention is paid to the use of explainable artificial intelligence (xAI) tools, in particular SHAP and LIME, to identify keywords in texts that most affect the model's decisions. A similar study was also conducted by the authors [10], however, coalitional game theory approaches were used here, which allowed us to analyze the contribution of each linguistic characteristic to the final evaluation of the text, as well as to derive a general linguistic profile of propaganda in the American media. Unlike the previous study, which investigated how vulnerable the models are to changes in critical words identified by xAI methods in order to assess their resistance to deliberate attacks, here the focus is on explaining the model's decisions through the interpretation of linguistic features that shape the propaganda message.

The authors of [11] emphasize that although modern artificial intelligence algorithms, in particular deep and machine learning methods, demonstrate high performance in many applied tasks, their opacity and tendency to biased decisions create serious ethical and practical challenges. These algorithms often operate as "black boxes", which makes it difficult to interpret the results, especially in the context of complex and sensitive tasks related to social discourse analysis. In this context, the potential of XAI is explored, which provides new tools for interpreting and explaining the decisions of machine learning models. The authors analyze XAI as a promising approach to increase the transparency of systems that detect destructive online content, in particular hate speech and disinformation.

In our own previous studies [12], we point out the importance of marker-oriented approaches, where the use of semantic features allows us to link certain linguistic structures with specific propaganda techniques. In such approaches, visual analytics plays an active role, which improves the interpretability of the results [13].

Despite the rapid development of artificial intelligence tools, the use of deep language models to detect propaganda is accompanied by a number of significant limitations. These include insufficient transparency of decisions, the risk of algorithmic bias, and the inconsistency of individual technical solutions with modern ethical and legal requirements. Existing approaches focus mainly on increasing the accuracy of classification or on studying the vulnerabilities of models to manipulative attacks, but they do not pay attention to the issue of ensuring the interpretability of the results in the context of compliance with the principles of digital justice, user rights protection, and regulatory soundness.

Therefore, based on the above analysis of existing solutions, the purpose of research is to substantiate the conceptual foundations and principles of implementing the multi-level approach to identifying propaganda techniques and objects of influence in text content, taking into account the requirements for transparency, explainability and responsibility of decision-making.

To achieve the set goal, the following research tasks must be performed:

1. To substantiate the architectural and conceptual principles of a multi-level approach to identifying propaganda techniques and objects of influence in text content, taking into account the principles of transparency, explainability and ethical responsibility.
2. To implement a model of primary classification of texts by the presence of signs of propaganda, using neural network technologies and a probability scale for differentiating messages by the degree of severity of manipulative influence.
3. To develop a methodology for identifying propaganda techniques at the level of semantic interpretation, using marker-oriented analysis and built-in means of visual interpretation of results.
4. To propose the approach to identifying objects of propaganda influence, which involves semantic grouping and establishing logical connections between rhetorical techniques and target concepts mentioned in the text.
5. To ensure transparency and explainability of model solutions by integrating local interpretation tools, as well as verify their effectiveness by comparing them with expert assessments.
6. To assess the effectiveness of the proposed system in real-world applications, in particular in the field of information security, to increase user trust and compliance with legal and ethical standards.

#### Proposed approach

The research proposes approach that implements multi-level processing of text content to identify propaganda techniques and corresponding objects of influence. The approach (Fig. 1) consists in decomposing the initial task of detecting a text containing propaganda, taking into account the requirements for transparency, explainability and responsibility of the decisions made, into successive tasks:

- (1) initial classification of the text for the presence or absence of signs of propaganda;
- (2) semantic interpretation of the techniques, with the identification of specific rhetorical or psychological techniques inherent in the propaganda discourse;
- (3) detection of objects of influence aimed at identifying the goals of propaganda influence and establishing their connection with the corresponding techniques.

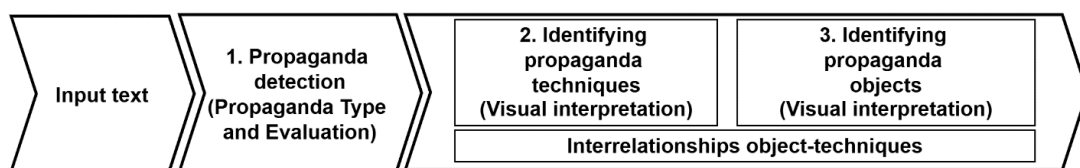


Fig. 1. General scheme of the approach to identifying propaganda techniques and objects

The input data of proposed approach is a text message.

Next, using pre-trained deep learning models, for the first task, using a binary classifier with a probable result, it is determined whether the text contains signs of propaganda, classifying messages by the level of probability into one of the categories: “non-propaganda”, “suspicious” or “propaganda”. If the probability of influence is detected, then we proceed to the second task.

The second task is responsible for detecting propaganda techniques and their visual interpretation. It is used only for texts classified as «propaganda text». The input text is fed in turn to 17 trained neural network models to analyze the presence of 17 propaganda techniques [14, 15, 16]:

1. «Appeal to fear-prejudice».
2. «Causal Oversimplification».
3. «Doubt».
4. «Exaggeration».
5. «Flag-Waving».
6. «Labeling».
7. «Loaded Language».

8. «Minimisation».
9. «Name Calling».
10. «Repetition».
11. «Appeal to Authority».
12. «Black and White Fallacy».
13. «Reductio ad hitlerum».
14. «Red Herring».
15. «Slogans».
16. «Thought terminating Cliches».
17. «Whataboutism».

Accordingly, the output data will be an assessment of the presence of propaganda techniques by markers [17] and a visual interpretation of the results [18].

The third task is responsible for detecting propaganda objects and their visual interpretation. It uses the results obtained during the implementation of previous methods. It transforms the input data into a set of thematic propaganda objects with the relationships of the detected objects with propaganda techniques.

Thus, the proposed approach not only ensures the detection of propaganda, but also meets the requirements for ethical responsibility: in particular, the transparency of models, the user's right to explain the results and the verifiability of the connections between influence techniques and the objects of their application.

### Results and discussion

For the proposed approach, an experimental research was conducted to assess the effectiveness of propaganda detection, classification of its techniques, and identification of influence objects, with an emphasis on transparency and interpretation of the obtained solutions.

Text classification by propaganda content, using a hybrid model based on BiLSTM [19] with an additional level of attention, showed on test datasets an F1-measure value of 0.91 when classifying Ukrainian-language texts included in the corpus of news and social messages. The value of the Recall metric (0.93) confirmed the model's ability to identify even weakly expressed forms of manipulative influence.

Recognition of propaganda techniques, using a model based on markers and the BERT architecture [20], showed effective differentiation between 17 classes of rhetorical techniques. F1-measure values in the range of 0.82–0.88 were obtained for the main categories («Appeal to Fear», «Loaded Language», «Name Calling»), which confirms the ability of the system to detect stable patterns even in stylistically heterogeneous texts.

When identifying objects of influence, the results confirmed the feasibility of combining the NER model [21] with semantic grouping mechanisms. Objects marked not only as named entities, but also through contextual associations (for example, mentions through pronouns, descriptive names, generalizations), were successfully associated with the corresponding propaganda techniques. This approach allowed us to avoid fragmentary analysis and provide a holistic picture of the connections between objects and rhetorical strategies.

Particular attention was paid to explaining the decisions made. The implemented LIME tools [22] provided an opportunity to illustrate which text fragments were key in determining the technique or object. An example of using local interpretation is shown in Fig. 2.

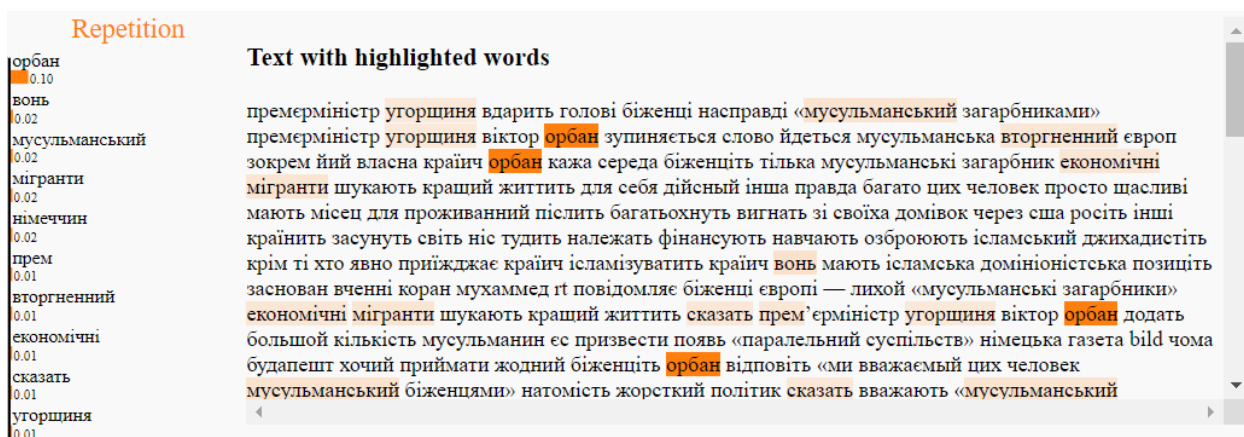


Fig.2. Using of local text interpretation

For example, in cases where the «Flag-Waving» technique was detected, tokens with high emotional modality and references to patriotic symbols gained the greatest weight, which was confirmed in LIME visualizations.

Expert evaluation of the explanations accompanying the automatic conclusions showed a high correlation between the interpretation of the model and the human perception of the influence structure. The consistency of the



results obtained automatically with the assessments of experts (Inter-Annotator Agreement at the level of 0.87) indicates the quality of not only the technical implementation, but also the ethical justification. An example of comparing the assessment of an expert with the developed software is shown in Fig.3.

**Analysis Result:**

The set of named entities with semantically close objects according to the analysis of contextual dependencies:

**ЗСУ**, ORG, вранці (0.21), вулиця (0.17), суперечка (0.17), виникнути (0.16)  
**донбас**, LOC, зма (0.26), випивати (0.22), разом (0.20), раніше (0.17)  
**донецький область**, LOC, вони (0.17), виникнути (0.15)

**Set of propaganda objects in the text:**

!! Смерть двох військовослужбовців **ЗСУ** в селі на **Донбасі**  
 На **Донбасі** один з військовослужбовців застрелився з табельної зброї, коли **вранці** знайшов поруч з собою мертвим свого товариша по службі.  
 За інформацією видання, майор і сержант **разом випивали** у житловому будинку, після чого між **ними виникла суперечка**, яка переросла у бійку на підвір'ї.  
 Майор сильно побив сержанта, що той втратив свідомість і замерз на **вулиці**. Вранці військовослужбовець виявив товариша по службі мертвим і застрелився з табельної зброї.  
 Як повідомлялося **раніше**, в **Донецькій області** у житловому будинку знайшли мертвими двох військових **ЗСУ**. За інформацією ЗМІ, у одного з **них**, майора, вогнепальне поранення голови, у другого — забій голови.

**Power of techniques used and their associated thematic objects:**

**The used techniques:**

1. Loaded Language. Expressed at 0.582
2. Repetition. Expressed at 0.317

Assessment of propagandistic objects belonging to the used techniques:

{**ЗСУ** (ORG) Added thematic set: [вранці, вулиця, суперечка, виникнути]} **Assessments of belonging:** [Loaded Language 0.593; Repetition 0.612]  
 {**донбас** (LOC) Added thematic set: [зма, випивати, разом, раніше]} **Assessments of belonging:** [Loaded Language 0.407; Repetition 0.35]  
 {**донецький область** (LOC) Added thematic set: [вони, виникнути]} **Assessments of belonging:** [Loaded Language 0.361; Repetition 0.71]

**Fig. 3. An example of comparing objects and techniques of propaganda**

The use of the «Charged Language» technique is used in the text to describe conflicts and violence, for example, «сильно побив», «замерз на вулиці», «застрелився з табельної зброї». This corresponds to the purpose of discrediting the Armed Forces and portraying them in a negative light, which corresponds to the expert's conclusion.

The use of the «Repetition» technique is used in the form of repeating information about the death of servicemen and violent actions. Repetition helps to enhance the negative impact and strengthen the negative impression. This corresponds to the purpose of persuading citizens not to join the ranks of the army, and active servicemen to resign.

According to the expert's conclusion (Fig.4), the theses about the alleged abuse of the Ukrainian military and the demoralization of servicemen were intended to: discredit the Armed Forces, the National Guard and other military formations in the eyes of Ukrainian citizens; convince Ukrainian citizens not to join the ranks of the Ukrainian army, and active servicemen to resign from its ranks.

Ці тези про нібито зловживання українських військових та деморалізацію військовослужбовців мали на меті:

- дискредитувати Збройні Сили, Національну Гвардію та інші військові формування в очах громадян України;
- переконати громадян України не вступати до лав українського війська, а діючих військовослужбовців – звільнитись із його лав;
- надати важливості Telegram-каналу «НачШтабу» як джерелу нібито унікальних новин, про які «не розповість» військове командування України, щоб завоювати довіру військовослужбовців-підписників, з подальшою метою спонукати їх ділитись інформацією, зокрема, службового і таємного характеру;

Окремою складовою кампанії із «висвітлення морального занепаду ЗСУ» є аномально велика кількість повідомлень «НачШтабу» про самогубства військовослужбовців. За 2019 рік було зафіксовано щонайменше 38 повідомлень, в 2020 – щонайменше 19, а в 2021 – 20, у яких не було наведено жодних доказів, що ці історії справді мали місце і в тому вигляді, в якому це подавав Telegram-канал.

**Fig. 4. Analysis of a text containing propaganda («Center of Strategic Communications» [23])**

Overall, the experimental results confirm that the combination of neural network technologies with built-in transparency mechanisms allows for the creation of systems that can be used not only as an analysis tool, but also as

an object of public trust. The application of the system in real conditions – in particular, in the activities of cyber police units and public organizations – revealed its practical value both in terms of efficiency and ethical responsibility.

### Conclusions

The conducted research shows that the combination of neural network technologies with methods of explainable artificial intelligent text processing allows creating an effective and at the same time ethically balanced system for detecting propaganda techniques in natural language messages. The proposed architecture provides not only high accuracy of propaganda content classification, but also demonstrates the ability to identify objects of influence and establish semantic connections between rhetorical strategies and target concepts.

The results of local interpretation and comparison with expert assessments have demonstrated the relevance and reliability of the system in the context of ethical responsibility. At the same time, the explainability of decisions, visualization of influential text fragments and transparency of computational procedures have become the basis for increasing trust in such technologies both from the user and from regulatory authorities.

In summary, it can be argued that the combination of technical efficiency with legal and ethical guarantees forms a new paradigm of responsible artificial intelligence, capable not only of detecting information threats, but also of functioning within the framework of socially acceptable and legally correct practices.

### References

1. Miller K. M., Lukic K., Skiera B. The impact of the General Data Protection Regulation (GDPR) on online tracking. *International Journal of Research in Marketing*. 2025. URL: <https://doi.org/10.1016/j.ijresmar.2025.03.002> (date of access: 05.06.2025).
2. Van Kolfchooten H., Van Oirschot J. The EU Artificial Intelligence Act (2024): Implications for healthcare. *Health Policy*. 2024. Vol. 149. P. 105152. URL: <https://doi.org/10.1016/j.healthpol.2024.105152> (date of access: 05.06.2025).
3. Fjeld J. et al. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI [Electronic resource] / J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, M. Srikumar. – Cambridge, MA: Berkman Klein Center for Internet & Society, 2020. – 121 p. – Mode of access: <https://www.abaj.ai/doc/papers/fjeld2020.pdf>. – Title from screen. – Accessed: 02 June 2025.
4. UNESCO. UNESCO's Recommendation on the Ethics of Artificial Intelligence: Key Facts. – Paris: United Nations Educational, Scientific and Cultural Organization, 2023. – 16 p.
5. Kazim E., Soares Koshiyama A. Human Centric AI: A Comment on the IEEE's Ethically Aligned Design. *SSRN Electronic Journal*. 2020. URL: <https://doi.org/10.2139/ssrn.3575140> (date of access: 06.06.2025).
6. Розпорядження Кабінету Міністрів України від 16 лютого 2024 р. № 137-р «Про затвердження плану пріоритетних дій Уряду на 2024 рік», Кабінет Міністрів України. URL: <https://www.kmu.gov.ua/npas/pro-zatverdzhennia-planu-priorytetnykh-dii-uriadu-na-2024-rik-137r-160224> (дата звернення: 13.03.2025).
7. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles / G. Da San Martino et al. Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online). Stroudsburg, PA, USA, 2020. URL: <https://doi.org/10.18653/v1/2020.semeval-1.186> (дата звернення: 13.04.2025).
8. Overview of the WANLP 2022 Shared Task on Propaganda Detection in Arabic / F. Alam et al. Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), Abu Dhabi, United Arab Emirates (Hybrid). Stroudsburg, PA, USA, 2022. URL: <https://doi.org/10.18653/v1/2022.wanlp-1.11> (дата звернення: 13.04.2025).
9. Cavaliere D., Gallo M., Stanzione C. Propaganda Detection Robustness Through Adversarial Attacks Driven by eXplainable AI. *Communications in Computer and Information Science*. Cham, 2023. P. 405–419. URL: [https://doi.org/10.1007/978-3-031-44067-0\\_21](https://doi.org/10.1007/978-3-031-44067-0_21) (date of access: 06.06.2025).
10. Barfar A. A linguistic/game-theoretic approach to detection/explanation of propaganda. *Expert Systems with Applications*. 2022. Vol. 189. P. 116069. URL: <https://doi.org/10.1016/j.eswa.2021.116069> (date of access: 06.06.2025).
11. Gongane V. U., Munot M. V., Anuse A. D. A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*. 2024. URL: <https://doi.org/10.1007/s42001-024-00248-9> (date of access: 06.06.2025).
12. Method of Semantic Features Estimation for Political Propaganda Techniques Detection Using Transformer Neural Networks / I. Krak, M. Molchanova, V. Didur, O. Sobko, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3917, pp. 286–297. URL: <https://ceur-ws.org/Vol-3917/paper56.pdf> (дата звернення: 19.03.2025).
13. Method for Neural Network Detecting Propaganda Techniques by Markers With Visual Analytic / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, E. Manziuk, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3790, pp. 158–170. URL: <https://ceur-ws.org/Vol-3790/paper14.pdf> (дата звернення: 19.03.2025).
14. Fine-Grained Analysis of Propaganda in News Article / G. Da San Martino et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Stroudsburg, PA, USA, 2019. URL: <https://doi.org/10.18653/v1/d19-1565> (дата звернення: 13.04.2025).
15. Think Fast, Think Slow, Think Critical: Designing an Automated Propaganda Detection Tool / L. Zavolokina et al. CHI '24: CHI Conference on Human Factors in Computing Systems, Honolulu HI USA. New York, NY, USA, 2024. URL: <https://doi.org/10.1145/3613904.3642805> (дата звернення: 04.12.2024).
16. Chow W. M., Levin D. H. The Diplomacy of Whataboutism and US Foreign Policy Attitudes. *International Organization*. 2024. Vol. 78, no. 1. P. 103–133. URL: <https://doi.org/10.1017/s002081832400002x> (дата звернення: 13.04.2025).
17. Молчанова М.О., Бармак О.В. Метод інтелектуального виявлення технік пропаганди за ознаками з використанням машинного навчання. *Науковий журнал «Наукові праці Донецького національного технічного університету»*, серія «Проблеми моделювання та автоматизації проектування». 2025. №1 (21). С. 76–85. <https://doi.org/10.31474/2074-7888> (дата звернення: 19.03.2025).
18. Молчанова М. Метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень. *Науковий журнал «Вісник Хмельницького національного університету»* серія: Технічні науки. 2024. Т. 343, № 6(1). С. 179–185. URL: <https://doi.org/10.31891/2307-5732-2024-343-6-27> (дата звернення: 19.03.2025).
19. Merryton A. R., Gethsiyal Augusta M. An Attribute-wise Attention model with BiLSTM for an efficient Fake News Detection. *Multimedia Tools and Applications*. 2023. URL: <https://doi.org/10.1007/s11042-023-16824-6> (дата звернення: 02.04.2025).
20. Large Language Models: Comparing Gen 1 Models (GPT, BERT, T5 and More). *Dev*. URL: <https://dev.to/admantium/large-language-models-comparing-gen-1-models-gpt-bert-t5-and-more-74h> (дата звернення: 13.03.2025).
21. Wilkho R. S., Gharaibeh N. G. FF-NER: A named entity recognition model for harvesting web-based information about

flash floods and related infrastructure impacts. International Journal of Disaster Risk Reduction. 2025. Vol. 125. P. 105604. URL: <https://doi.org/10.1016/j.ijdr.2025.105604> (date of access: 06.06.2025).

22. ELI5.LIME: Explain PyTorch Text Classification Network Predictions Using LIME Algorithm, CoderzColumn, 2024. URL: <https://coderzcolumn.com/tutorials/artificial-intelligence/eli5-lime-explain-pytorch-text-classification-network-predictions> (дата звернення: 06.04.2025).

23. Центр стратегічних комунікацій, Spravdi, 2025. URL: <https://spravdi.gov.ua/> (дата звернення: 06.04.2025).

<b>Маруна Molchanova Марина Молчанова</b>	Postgraduate student, Department of Computer Science, Khmelnytskyi National University <a href="https://orcid.org/0000-0001-9810-936X">https://orcid.org/0000-0001-9810-936X</a> e-mail: <a href="mailto:m.o.molchanova@gmail.com">m.o.molchanova@gmail.com</a>	Аспірант кафедри комп'ютерних наук, Хмельницький національний університет
<b>Pawan Kumar Dutt Паван Кумар Датт</b>	Doctor of Philosophy, Senior Lecturer at the School of Law, Tallinn University of Technology (Estonia). <a href="https://orcid.org/0000-0001-8772-0315">https://orcid.org/0000-0001-8772-0315</a>	Доктор філософії, старший викладач Школи права Талліннського технічного університету (Естонія).