УДК 004.8:004.056:519.86
DOI https://doi.org/10.24144/2616-7700.2026.49(2).230-237

А. А. Матей

ДВНЗ «Ужгородський національний університет»,
аспірант кафедри програмного забезпечення систем
andrii.matei@uzhnu.edu.ua
ORCID: https://orcid.org/0009-0001-0280-1763

НЕЧІТКА МОДЕЛЬ РОЗПІЗНАВАННЯ ШТУЧНО СТВОРЕНОГО КОНТЕНТУ СОЦІАЛЬНИМИ КЛАСАМИ В КОНТЕКСТІ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ ДЕРЖАВИ

У статті розглянуто науково-методичні аспекти розроблення нечіткої моделі оцінювання рівня розпізнавання штучно створеного контенту різними соціальними класами в контексті інформаційної безпеки держави. Актуальність дослідження зумовлена стрімким поширенням генеративного штучного інтелекту, зростанням обсягів штучно створеного контенту та підвищенням ризиків його маніпулятивного впливу на суспільство. Проаналізовано сучасні наукові підходи до виявлення AI-згенерованого контенту та обґрунтовано наукову прогалину, що полягає у недостатньому врахуванні соціально-демографічних характеристик громадян, їхньої здатності до розпізнавання такого контенту та відсутності інтегральних моделей для підтримки управлінських рішень.

Запропоновано нечітку модель, яка базується на формалізації кількісних і якісних характеристик соціального профілю громадян, опрацюванні їхніх відповідей щодо визначення авторства текстових новин та використанні вектора профілю соціального класу, що формується особою, яка приймає рішення. Модель реалізує процедуру ранжування респондентів за ступенем близькості до заданого соціального класу та дає змогу визначати інтегровану оцінку рівня розпізнавання штучно створеного контенту. На основі отриманого значення здійснюється лінгвістична інтерпретація рівня технічного ризику маніпулятивного впливу штучно створеного контенту на інформаційну безпеку держави. Використання функцій належності, мір близькості, вагових коефіцієнтів і згорткових процедур забезпечує гнучкість, формалізованість, відтворюваність та інтерпретованість результатів оцінювання.

Практична значущість отриманих результатів полягає у можливості використання запропонованої моделі органами державного управління, структурами у сфері кібербезпеки, освітніми установами та аналітичними центрами для виявлення вразливих соціальних груп, оцінювання рівня їхньої стійкості до маніпулятивного впливу штучно створеного контенту та обґрунтування цільових превентивних заходів у сфері інформаційної безпеки держави.

Ключові слова: штучно створений контент, інформаційна безпека держави, соціальний клас, експертне оцінювання, нечітка модель, ризик, підтримка прийняття рішень.

1. Вступ. Стрімкий розвиток генеративного штучного інтелекту зумовив значне зростання обсягів штучно створеного контенту, який дедалі складніше відрізнити від створеного людиною. Це формує нові виклики для інформаційної безпеки держави, оскільки такі технології можуть використовуватися для маніпулятивного впливу, поширення дезінформації та формування викривлених інформаційних наративів.

У сучасних дослідженнях основну увагу приділено технічним методам виявлення такого контенту, однак недостатньо вивченим залишається людський фактор — здатність громадян до його розпізнавання та рівень вразливості до

маніпулятивного впливу. У зв'язку з цим виникає потреба в розробленні підходів, що дають змогу формалізувати й оцінювати рівень розпізнавання штучно створеного контенту з урахуванням соціально-демографічних характеристик населення.

Для розв'язання цієї задачі доцільним є використання апарату нечітких множин і нечіткої логіки, що дозволяє враховувати невизначеність, суб'єктивність оцінок і нечіткість меж між соціальними групами. Такий підхід забезпечує інтеграцію кількісних і якісних характеристик у межах єдиної моделі.

Метою статті є розроблення нечіткої моделі оцінювання рівня розпізнавання штучно створеного контенту різними соціальними класами в контексті технічних ризиків інформаційної безпеки держави.

2. Огляд літератури. Упродовж останніх років проблема штучно створеного контенту набула особливої актуальності через стрімкий розвиток генеративного штучного інтелекту. Сучасні дослідження зосереджені на питаннях сприйняття дезінформації, індивідуальної вразливості до неї та чинниках довіри до хибних повідомлень. Зокрема, показано, що люди часто спираються на інтуїтивні, але помилкові евристики під час розпізнавання штучно створеного тексту [1–2]. Це свідчить, що проблема має не лише технічний, а й когнітивно-соціальний вимір.

Окремий напрям досліджень стосується емпіричної оцінки здатності людей розпізнавати штучно створений контент у різних модальностях. Узагальнення сучасних праць показує, що точність виявлення deepfake-контенту часто є низькою, особливо без спеціальної підготовки, а складність розпізнавання підтверджується для текстових, аудіо- та відеоформатів [3–5]. Це свідчить, що ризик маніпулятивного впливу зумовлений не лише якістю генерації контенту, а й обмеженими можливостями людини щодо його ідентифікації.

Водночас сучасні дослідження підтверджують вплив соціальних, поведінкових і психологічних чинників на сприйняття штучно створеного контенту. Показано, що точність його розпізнавання залежить від індивідуальних когнітивних характеристик, цифрових звичок і особливостей аудиторії, тоді як маркування AI-згенерованого контенту має лише частковий ефект [6–8]. Крім того, AI-згенеровані повідомлення можуть бути майже такими ж або навіть більш переконливими, ніж створені людиною, зокрема у суспільно значущих темах [9–10]. Це посилює потребу оцінювати не лише технічну детекцію, а й соціальну вразливість різних груп населення.

Таким чином у науковій літературі бракує комплексних моделей, які б поєднували соціально-демографічні характеристики громадян, результати розпізнавання такого контенту та інтерпретовану оцінку ризику для інформаційної безпеки держави.

3. Матеріали та методи. Нехай розглядається множина громадян $C = \{c_1; c_2; \ldots; c_n\}$, які підлягають оцінюванню за рівнем розпізнавання ними штучно створеного контенту. Громадяни досліджуються за регіоном проживання або за розміром населеного пункту проживання. Кожен такий суб'єкт дослідження характеризується множиною ознак $S = \{s_1; s_2; \ldots; s_m\}$, причому $f : C \to S$. Очевидно, що кожен громадянин має власний соціальний профіль, який визначається сукупністю цих характеристик. Групування характеристик множини S за певними правилами $T = \{t_1; t_2; \ldots; t_u\}$ дає змогу сформувати

різні соціальні класи $CL(T) = \{cl_1(t_1); cl_2(t_2); \ldots; cl_k(t_u)\}$, де $CL(T) \subseteq S$.

Для громадян пропонується деякий набір текстових новин $R = \{R_1; R_2; \ldots; R_l\}$ з метою визначення їх авторства, тобто встановлення, чи написаний текст людиною, чи створений штучним інтелектом. Сукупність показників для оцінювання характеристик громадян, їхніх відповідей щодо розпізнавання тексту, а також система опрацювання вхідних даних формують інформаційну модель оцінювання рівня розпізнавання штучно створеного контенту громадянами — K_{TR}. Опрацьовані вхідні дані надалі використовуються в нечіткій моделі розпізнавання штучно створеного контенту соціальними класами в контексті інформаційної безпеки держави — M_{TR}.

Формально нечітка модель може бути представлена у вигляді оператора:

$$M_{TR}(C,\ S,\ T,\ CL,\ K_{TR}) \to f_{TR}(m_{TR},\ TR). \tag{1}$$

Відповідно до вхідних змінних C, S, T, CL, K_{TR} визначається множина вихідних значень f. Результатом функціонування нечіткої моделі є: m_{TR} — інтегрована оцінка рівня розпізнавання штучно створеного контенту для відповідного соціального класу CL; TR — лінгвістична оцінка рівня технічного ризику маніпулятивного впливу штучно створеного контенту на інформаційну безпеку держави. Для ілюстрації нечіткої моделі представляється структурна схема, рис. 1.

Нехай кожен громадянин C описується множиною характеристик соціального профілю

$$S = \{S_1(su_1); S_2(su_2); \ldots; S_m(su_m)\}, \quad f : C \to S. \tag{2}$$

Характеристики можуть мати кількісну або якісну природу, тому підлягають формалізації й нормуванню. В інформаційній моделі використано такі характеристики: стать S_1, вік S_2, рівень освіти S_3, дохід родини S_4, а також частотні показники користування інформаційними каналами S_5–S_8. Їх лінгвістичні оцінки переводяться у формалізований вигляд за допомогою характеристичних функцій μ_1–μ_8, зокрема на основі дискретних шкал і терм-множин.

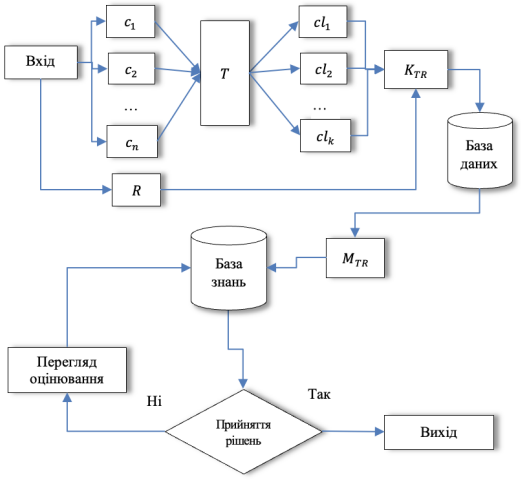

Рис. 1. Структурна схема нечіткої моделі.

Формально перехід від лінгвістичної оцінки x_j до нормованого вигляду задається характеристичною функцією $\mu_j(x_j) : X_j \to [0,1]$, де X_j — область значень j-ї соціальної характеристики, а $\mu_j(x_j)$ визначає ступінь належності конкретного значення до відповідного терму або інтервалу. Для дискретних характеристик $\mu_j(x_j) \in \{0,1\}$, тоді як для порядкових та інтервальних характеристик використовуються кусочно-задані або трапецієподібні функції належності, що забезпечують уніфіковане нормування вхідних даних для подальших обчислень у моделі.

Для характеристик S_5–S_8, що відображають інтенсивність користування інформаційними каналами, застосовується єдина шкала частотності — від «ніколи» до «більше 6 годин на добу», що забезпечує уніфіковане подання вхідних параметрів.

Для оцінювання рівня розпізнавання штучно створеного тексту респондентам пропонується набір новин R, для кожної з яких потрібно визначити авторство двох текстів: $t_h(\alpha)$, створеного людиною, та $t_h(\beta)$, згенерованого штучним інтелектом. Відповіді формалізуються термами $r = \{\alpha, \beta\}$, а правильність розпізнавання авторства описується функціями $\tau_{ih1}(\alpha)$ та $\tau_{ih2}(\beta)$.

Для r_{ih1} маємо:

$$\tau_{ih1}(\alpha) = \begin{cases} 0, & \text{якщо для новини } t_h(\alpha) \text{ відповідь } r_{ih1} = \beta; \\ 1, & \text{якщо для новини } t_h(\alpha) \text{ відповідь } r_{ih1} = \alpha. \end{cases} \tag{3}$$

Для r_{ih2} маємо:

$$\tau_{ih2}(\beta) = \begin{cases} 0, & \text{якщо для новини } t_h(\beta) \text{ відповідь } r_{ih2} = \alpha; \\ 1, & \text{якщо для новини } t_h(\beta) \text{ відповідь } r_{ih2} = \beta. \end{cases} \tag{4}$$

У нечіткій моделі M_{TR}формалізовані оцінки використовуються лише для обчислень і не характеризують особисті якості громадян. Їх призначення полягає у визначенні рівня розпізнавання штучно створеного контенту різними соціальними класами. На першому етапі моделі задаються інформаційні показники характеристик громадян і формалізується підхід до опрацювання відповідей щодо авторства текстових новин.

На другому етапі опрацьовані вхідні дані обчислюються за нечіткою моделлю — M_{TR}.

Нехай розглядається множина громадян C, які пройшли оцінювання згідно з інформаційною моделлю K_{TR}. На першому етапі для кожного респондента вже отримано нормовані оцінки $\mu_1(su_1), \mu_2(su_2), \ldots, \mu_m(su_m)$ за характеристиками соціального профілю S. Крім того, для множини новин R отримано формалізовані відповіді респондентів $\tau_{ih1}(\alpha)$ та $\tau_{ih2}(\beta)$ щодо розпізнавання авторства двох текстів: створеного людиною $t_h(\alpha)$ та штучним інтелектом $t_h(\beta)$.

На першому кроці другого етапу для кожного громадянина визначається нормована оцінка рівня розпізнавання штучно створеного контенту по кожній новині:

$$\lambda_{ih} = \frac{1}{2}(\tau_{ih1}(\alpha) + \tau_{ih2}(\beta)), \quad i = \overline{1,n}; \quad h = \overline{1,l}. \tag{5}$$

Таким чином, отримуються значення $\lambda_{ih} \in [0;1]$, де більші значення відповідають вищій точності розпізнавання.

Формула (5) у наведеному дослідженні відповідає нейтральному випадку однакової ваги двох типів помилок. За потреби її можна модифікувати шляхом введення ваг ω_1 і ω_2, причому для задач інформаційної безпеки доцільно задавати $\omega_2 > \omega_1$, якщо пропуск штучно створеного маніпулятивного контенту є більш небезпечним.

Далі для кожного респондента обчислюється узагальнена оцінка рівня розпізнавання штучно створеного контенту:

$$\omega_i = \frac{1}{l} \sum_{h=1}^{l} \lambda_{ih}. \tag{6}$$

Оскільки дослідження орієнтоване на соціальні класи, характеристики S групуються за правилом T, унаслідок чого формується соціальний клас $cl_d(t_c)$, заданий вектором профільних вимог. Нормовані характеристики респондентів подаються матрицею рішень M, а заданий ОПР вектор профілю після застосування характеристичних функцій перетворюється у вектор кількісних змінних Z. На цій основі визначаються відносні оцінки близькості характеристик громадян до профілю відповідного соціального класу:

$$q_{di} = 1 - \frac{\mid z_d - \mu_{di}(su_d) \mid}{\max\{z_d - \min\limits_i(\mu_{di}(su_d)); \max\limits_i(\mu_{di}(su_d)) - z_d\}}. \tag{7}$$

Матриця $Q = \{q_{di}\}$ відображає ступінь близькості кожного громадянина до профілю відповідного соціального класу.

За потреби ОПР задає вагові коефіцієнти $\{v_1, v_2, \ldots, v_k\}$ для характеристик соціального профілю. На їх основі обчислюються нормовані ваги:

$$w_d = \frac{v_d}{\sum\limits_{d=1}^{k} v_d}; \quad w_d \in [0; 1]. \tag{8}$$

При цьому виконується умова $\sum\limits_{d=1}^{k} w_d = 1$.

Зауважується, що початкові вагові коефіцієнти v_d рекомендується визначати методом попарних порівнянь Сааті. У цьому разі v_d обчислюються як компоненти нормованого власного вектора матриці попарних порівнянь, сформованої особою, що приймає рішення.

Далі будується ранжувальний ряд громадян відносно вектора профілю соціального класу за допомогою середньої згортки:

$$m(c_i) = \sum_{d=1}^{k} w_d \cdot q_{di}. \tag{9}$$

На основі отриманого ранжування ОПР задає кількість громадян g, $g < n$, які найбільше відповідають профілю вибраного соціального класу. Для цих g громадян за узагальненими оцінками ω_c визначається інтегрована оцінка:

$$m_{TR} = \frac{1}{g} \sum_{c=1}^{g} \omega_c. \tag{10}$$

На завершальному етапі значення $m_{TR} \in [0;1]$ інтерпретуються лінгвістичною оцінкою рівня технічного ризику маніпулятивного впливу штучно створеного контенту на інформаційну безпеку держави за наступною терм-множиною $TR = \{tr_1, tr_2, \ldots, tr_5\}$. При цьому: $m_{TR} \in (0.8;1]$ — tr_1: дуже низький рівень; $m_{TR} \in (0.6;0.8]$ — tr_2: низький рівень; $m_{TR} \in (0.4;0.6]$ — tr_3: середній рівень; $m_{TR} \in (0.2;0.4]$ — tr_4: високий рівень; $m_{TR} \in [0;0.2]$ — tr_5: дуже високий рівень.

Межі між термами необхідно визначити на основі реальних даних респондентів, яким запропонувати тексти для розпізнавання штучно створеного контенту.

4. Висновки та перспективи подальших досліджень. У дослідженні розроблено нечітку модель оцінювання рівня розпізнавання штучно створеного контенту різними соціальними класами в контексті інформаційної безпеки держави. Модель, на відміну від суто технічних підходів, враховує людський фактор і соціальний профіль респондентів, забезпечуючи отримання інтегрованої оцінки m_{TR} та лінгвістичної оцінки технічного ризику TR.

Наукова новизна підходу полягає у формалізації зв'язку між соціально-демографічними характеристиками, результатами розпізнавання штучно створених текстів і рівнем технічного ризику. Практична цінність моделі полягає у виявленні вразливих соціальних груп для обґрунтування превентивних рішень у сфері інформаційної безпеки.

Водночас підхід потребує подальшої емпіричної верифікації, калібрування меж лінгвістичних термів ризику та розширення множини вхідних характеристик. У подальших дослідженнях планується подати модельний числовий приклад покрокового розрахунку для демонстрації працездатності та відтворюваності моделі.

Конфлікт інтересів

Автор заявляє про відсутність конфлікту інтересів.

Фінансування

Дослідження було проведено без фінансової підтримки.

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автор підтверджує, що при створенні даної роботи він не використовув технології штучного інтелекту.

Список використаної літератури

1. Nan, X., Wang, Y., & Thier, K. (2022). Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine,* 314, 115398. https://doi.org/10.1016/j.socscimed.2022.115398
2. Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences,* 120(11), e2208839120. https://doi.org/10.1073/pnas.2208839120
3. Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports,* 16, 100538. https://doi.org/10.1016/j.chbr.2024.100538
4. Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE,* 18(8), e0285333. https://doi.org/10.1371/journal.pone.0285333
5. Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. W. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications,* 15(1), 7629. https://doi.org/10.1038/s41467-024-51998-z
6. Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports,* 14, 25989. https://doi.org/10.1038/s41598-024-76218-y
7. Lovato, J., St-Onge, J., Harp, R., Salazar Lopez, G., Rogers, S. P., Ul Haq, I., Hébert-Dufresne, L., & Onaolapo, J. (2024). Diverse misinformation: Impacts of human biases on detection of deepfakes on networks. *npj Complexity,* 1, 5. https://doi.org/10.1038/s44260-024-00006-y
8. Li, F., & Yang, Y. (2024). Impact of artificial intelligence–generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based, randomized, controlled experiment. *JMIR Formative Research,* 8, e60024. https://doi.org/10.2196/60024
9. Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus,* 3(2), pgae034. https://doi.org/10.1093/pnasnexus/pgae034
10. Salvi, F., Horta Ribeiro, M., Gallotti, R., & et al. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour,* 9, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Matei A. A. A fuzzy model of recognition of artificially created content by social classes in the context of state information security.

The article considers the scientific and methodological aspects of developing a fuzzy model for assessing the level of recognition of artificially created content by different social classes in the context of state information security. The relevance of the study is due to the rapid spread of generative artificial intelligence, the growth of the volume of artificially created content, and the increase in the risks of its manipulative impact on society. Modern scientific approaches to detecting AI-generated content are analyzed, and the scientific gap is substantiated, which consists of insufficient consideration of the socio-demographic characteristics of citizens, their ability to recognize such content, and the lack of integral models to support management decisions.

A fuzzy model is proposed, which is based on the formalization of quantitative and qualitative characteristics of the social profile of citizens, processing their answers regarding

the determination of the authorship of text news, and using the vector of the social class profile formed by the decision-maker. The model implements the procedure for ranking respondents by the degree of proximity to a given social class and allows for determining an integrated assessment of the level of recognition of artificially created content. Based on the obtained value, a linguistic interpretation of the level of technical risk of the manipulative influence of artificially created content on the information security of the state is carried out. The use of membership functions, proximity measures, weighting coefficients, and convolutional procedures ensures flexibility, formalization, reproducibility, and interpretability of the assessment results.

The practical significance of the results obtained lies in the possibility of using the proposed model by state administration bodies, structures in the field of cybersecurity, educational institutions, and analytical centers to identify vulnerable social groups, assess the level of their resistance to the manipulative influence of artificially created content, and justify targeted preventive measures in the field of information security of the state.

Keywords: artificially created content, information security of the state, social class, expert assessment, fuzzy model, risk, decision-making support.

Отримано: 20.03.2026 Прийнято: 07.04.2026 Опубліковано: 30.04.2026