

**Громенко Владислав Віталійович**

аспірант

Інститут програмних систем НАН України, Київ, Україна

ORCID: 0009-0001-5285-9912

gromvlad12@gmail.com

Шевченко Віктор Леонідович

доктор технічних наук, професор, заступник директора інституту з наукової роботи

Інститут програмних систем НАН України, Київ, Україна

професор кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна

ORCID: 0000-0002-9457-7454

gii2014@ukr.net

**ВИКОРИСТАННЯ ІНДЕКСУ МАНІПУЛЯТИВНОЇ НАСИЧЕНОСТІ ДЖЕРЕЛА В
БАГАТОКРИТЕРІАЛЬНОМУ ВІДБОРІ НОВИННИХ ПОВІДОМЛЕНЬ**

Анотація. Системи відбору новин, що ранжують повідомлення за показниками залучення, схильні підсилювати маніпулятивний зміст, причому маніпулятивність може бути сталою характеристикою редакційної практики окремих джерел. Відомі підходи зазвичай розглядають ці чинники окремо: вони або штрафують маніпулятивне повідомлення однаково незалежно від джерела, або понижують джерело загалом разом з його неманіпулятивними повідомленнями. Метою статті є дослідження адитивного й мультиплікативного способів урахування маніпулятивної насиченості джерела в багатокритеріальному відборі новинних повідомлень та їхнє порівняння на реальних даних. Маніпулятивність повідомлення оцінює класифікатор із кількома мітками за десятьма маніпулятивними техніками, побудований на великій мовній моделі. Індекс маніпулятивної насиченості джерела визначено як середню маніпулятивність повідомлень джерела в часовому вікні й перевірено як вимірювальний засіб. В адитивній формі індекс входить до цільової функції окремим доданком і понижує всі повідомлення джерела, у мультиплікативній – множить штраф за маніпулятивність повідомлення й не понижує неманіпулятивних повідомлень. Обидві форми порівняно зі штрафом без індексу й між собою за однакової ціни в релевантності на замороженому корпусі зі 125 987 повідомлень 34 джерел за 182 доби для 2 184 тематичних запитів. Штраф за маніпулятивність самого повідомлення знижує середню маніпулятивність показаних повідомлень у 5,4 раза ціною 1,09 % релевантності. За тієї самої ціни адитивна форма сильніше зміщує позиції від насичених джерел, але витісняє й їхні неманіпулятивні повідомлення; мультиплікативна форма майже не зачіпає неманіпулятивних повідомлень, проте зменшує частку позицій насичених джерел помірніше: на 15,6 % проти 62,8 % в адитивної відносно штрафу без індексу тієї самої ціни.

Ключові слова: машинне навчання; інформаційне середовище; новинні повідомлення; маніпулятивні техніки; виявлення маніпуляцій; великі мовні моделі; рекомендаційні системи; ранжування; багатокритеріальна оптимізація.

ВСТУП

Постановка проблеми. Системи автоматизованого відбору визначають, які повідомлення дістануться читачеві й у якому порядку. Кожне додаткове морально-емоційне слово в дописах на поляризовані теми збільшує їх поширення [1]. Аудит алгоритму стрічки, що ранжує дописи за прогнозованим залученням, показав, що порівняно з хронологічною стрічкою цей алгоритм підсилює показ емоційно забарвленого змісту, ворожого до зовнішньої групи, хоча самі користувачі таким політичним дописам переваги не віддають [2]. Оскільки маніпулятивні техніки «навантажена мова», «апеляція до страху», «страх, непевність та сумніви» спираються на емоційно забарвлену лексику, цільова функція, побудована на залученні, схильна винагороджувати маніпулятивні формулювання. Стаття 34 Регламенту Європейського Союзу про цифрові послуги зобов'язує дуже великі онлайн-платформи



оцінювати системні ризики, зокрема внесок у них рекомендаційних систем, з урахуванням регіональних і мовних аспектів [3], що надає практичного значення оцінюванню впливу відбору на інформаційне середовище.

У публікації «Карусель емоцій» видання Texty.org.ua, що охопила сто найпопулярніших публічних телеграм-каналів категорій «Новини і медіа» та «Політика», частка дописів щонайменше з однією маніпулятивною технікою різнилася між каналами: від менш як 2 % в одному до 90 % в іншому [4], тобто маніпулятивність розподілена між джерелами нерівномірно. Отже, однаково маніпулятивне повідомлення може надійти з потоку, де маніпуляції поодинокі, або з потоку, насиченого ними. Потік повідомлень джерела за часовий проміжок надалі називається середовищем джерела. Для виявлення маніпулятивних технік у текстах українською та російською мовами створено спеціалізований набір даних із бенчмарком [5], на якому можна навчити класифікатор маніпулятивності окремого повідомлення. Тому постає питання, як поєднати оцінку маніпулятивності повідомлення зі станом середовища джерела, з якого повідомлення походить, у цільовій функції відбору.

Аналіз останніх досліджень і публікацій. Helberger [6] обґрунтувала, що новинні рекомендаційні системи виконують суспільну функцію й мають оцінюватися за нормативними критеріями, а не лише за короткостроковими показниками на кшталт кількості переходів. Vrijenhoek та ін. [7], [8] перетворили цю позицію на обчислювані міри різноманіття показу – калібрування, фрагментації, репрезентативності, активації та альтернативних голосів. Ці міри описують розподіл показу за темами, поглядами й голосами і застосовуються до вже сформованого переліку; спосіб викладу окремого повідомлення вони не враховують.

Ближчими до задачі статті є підходи, що враховують у відборі властивість змісту окремого повідомлення. Jia та ін. [9] в онлайн-експерименті з макетом стрічки оцінювали антидемократичні погляди в дописах, зокрема великою мовною моделлю, і знижували такі дописи; міжпартійна ворожість учасників зменшилася без погіршення досвіду користування й залучення. Piccardi та ін. [10] у польовому експерименті показали, що переранжування стрічки за подібною оцінкою змінює афективну поляризацію користувачів. В обох роботах штраф не залежить від середовища джерела, а знижується чи вилучається лише саме повідомлення. Як сигнал ранжування запропоновано також оцінку якості джерела за судженнями неспеціалістів [11]; вона стосується джерела загалом, тож неманіпулятивне повідомлення джерела з низькою оцінкою знижується нарівні з маніпулятивним.

Спільне завдання UNLP 2025 [5] відкрило можливість автоматично досліджувати маніпулятивність у телеграм-каналах: воно надало набір дописів українського сегмента Telegram, розмічених за десятьма маніпулятивними техніками, і бенчмарк для порівняння класифікаторів, найкращі з яких досягли масо- F_1 близько 0,49 [5]. Класифікатор, навчений на цьому наборі, оцінює кожне окреме повідомлення, а отже, дає змогу оцінювати маніпулятивність джерела не одноразово, а в кожен момент за доступною історією його повідомлень. Природною основою для такої оцінки є усереднення за часовим вікном. Якщо ширину вікна задавати не сталою, а відстанню до k -го найближчого спостереження, кількість повідомлень у вікні не залежатиме від частоти публікацій джерела; так побудовано оцінки за найближчими сусідами [12].

Невирішеним лишається питання, як саме враховувати стан середовища джерела в цільовій функції відбору – додавати оцінку джерела до штрафу за маніпулятивність повідомлення чи масштабувати нею цей штраф – і як кожен зі способів змінює склад показу на реальних даних.

Мета статті. Метою статті є дослідження адитивного й мультиплікативного способів урахування маніпулятивної насиченості джерела в багатокритеріальному відборі новинних повідомлень та їхнє порівняння на реальних даних.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Оцінювання маніпулятивності повідомлення

Маніпулятивність повідомлення оцінює класифікатор із кількома мітками за десятьма маніпулятивними техніками таксономії спільного завдання UNLP 2025 [5] – «навантажена мова», «блискучі узагальнення», «ейфорія», «апеляція до страху», «страх, непевність та сумніви», «апеляція до народу», «кліше без суті», «Whataboutism», «вибіркова правда» й «опудало». Для кожної техніки j він видає оцінку ймовірності її наявності $p_j(a)$, а зведений показник маніпулятивності повідомлення (1) – середнє цих оцінок:

$$M(a) = \frac{1}{10} \sum_{j=1}^{10} p_j(a) \in [0, 1], \quad (1)$$

де a – повідомлення; $p_j(a)$ – оцінка ймовірності наявності техніки j у повідомленні a ; $M(a)$ –



показник маніпулятивності повідомлення.

Показник (1) залежить лише від тексту повідомлення, без даних про джерело чи сусідні повідомлення, тому індекс насиченості, побудований на $M(a)$, на вхід класифікатора не повертається. Класифікатор побудовано на великій мовній моделі, адаптованій до української мови [13], і навчено на навчальній частині набору UNLP 2025; на його випробувальній вибірці $\text{macro-}F_1$ становить 0,446.

Індекс маніпулятивної насиченості джерела

Корпус експерименту складається з повідомлень 34 публічних українськомовних телеграм-каналів, зібраних за 182 доби, з 15 лютого до 16 серпня 2026 року. Канали дібрано з переліку публікації «Карусель емоцій» видання Textu.org.ua [4], автори якої згрупували їх у п'ять категорій і впорядкували всередині кожної за рівнем маніпулятивності (далі – зовнішній рейтинг): 19 каналів обрано рівномірно за місцем у зовнішньому рейтингу, ще 15 – серед каналів з аудиторією від 500 тис. підписників. Загалом зібрано 255 362 повідомлення; індекс обчислюється за 125 987 повідомленнями від ста знаків.

Середовище джерела – потік саме його повідомлень, а не повідомлень, які спожив конкретний читач. Маніпулятивна насиченість характеризує редакційну практику джерела, тож її можна вимірювати без даних про користувачів і зіставляти між джерелами, а те, що вона змінюється повільно, перевірено в табл. 1. Повідомлення джерела у вікні задає множина (2):

$$H_c(t, \Delta) = \{a: \text{src}(a) = c, \quad t - \Delta \leq \tau(a) \leq t\}, \quad (2)$$

де $H_c(t, \Delta)$ – повідомлення джерела c у вікні шириною Δ , що завершується моментом t ; $\text{src}(a)$ – джерело повідомлення a ; $\tau(a)$ – момент його публікації.

Кількість повідомлень у вікні сталої ширини пропорційна частоті публікацій джерела: для джерела, що публікує рідко, оцінка спиралася б на кілька повідомлень, а її похибка різнилася б між джерелами майже в п'ятдесят разів (табл. 1). Тому вікно розширюється, доки не охопить щонайменше $n_{\min}$ повідомлень (3):

$$\Delta_c^*(t) = \min\{\Delta \geq \Delta_0: |H_c(t, \Delta)| \geq n_{\min}\}, \quad (3)$$

де $\Delta_c^*(t)$ – ширина вікна для джерела c у момент t ; $\Delta_0 = 24$ год – базова ширина вікна; $n_{\min} = 10$ – мінімальна кількість повідомлень у вікні, єдине з перевірених значень (10, 30, 50), за якого індекс визначений для всіх 34 джерел; $|\cdot|$ – кількість повідомлень у множині.

Найбільша ширина вікна – 30 діб; якщо й за неї вікно не наповнилося, індекс у момент t не визначений. На відміну від вікон оцінок за найближчими сусідами [12], вікно (3) однобічне – охоплює лише минуле – і має нижню межу Δ_0 . Індекс маніпулятивної насиченості джерела (4) – середній показник маніпулятивності повідомлень вікна:

$$P_c(t) = \frac{1}{|H_c(t, \Delta_c^*(t))|} \sum_{a \in H_c(t, \Delta_c^*(t))} M(a) \in [0, 1], \quad (4)$$

де $P_c(t)$ – індекс джерела c у момент t ; $M(a)$ – показник (1).

Індекс перевірено як вимірювальний засіб (табл. 1): чи розрізняє він джерела, чи стійкий до параметрів обчислення [14], чи відтворюється на незалежних половинах даних [15], чи узгоджується з незалежним вимірюванням [14] і чи зіставна його точність між джерелами. Вимірювання виконано в 182 добові моменти. Індекс визначений у 95,6 % пар «джерело – момент» і для всіх 34 джерел, вікно розширювалося в 33,1 % пар, а медіанна ширина вікна становить добу.

Таблиця 1

Перевірка індексу (4) за $n_{\min} = 10$

Критерій	Результат
Розрізнявальна здатність	стандартне відхилення між джерелами 0,0745, у часі 0,0214, відношення 3,48; середні крайніх джерел різняться в 38 разів
Стійкість до параметрів обчислення	120 сполучень п'яти параметрів (Δ_0 , $n_{\min}$, спосіб усереднення, поріг довжини, вилучення дослівних повторів); найменша рангова кореляція з базовим упорядкуванням – 0,965
Відтворюваність на незалежних половинах даних	5 909 пар «джерело – момент»; рангова кореляція 0,9924, з поправкою Спірмена – Брауна 0,9962
Узгодженість із незалежним вимірюванням	рангова кореляція із зовнішнім рейтингом у межах його категорій $-1,000$, $-0,821$ і $-0,759$; об'єднане $p = 1,2 \cdot 10^{-8}$
Зіставність точності між джерелами	відношення найбільшої серед джерел похибки індексу до найменшої: 23,2 за правила (3); 44,9–48,6 за вікон сталої ширини 24, 48 і 168 год



Мінливість між джерелами понад утричі перевищує мінливість того самого джерела в часі, тобто індекс вимірює сталу властивість джерела, а не поточну подієву ситуацію.

Зовнішній рейтинг упорядковує джерела лише в межах своїх категорій, тому узгодженість перевірено окремо в кожній категорії з достатньою кількістю джерел, а ймовірності випадкового збігу, обчислені перестановним тестом, об'єднано за Фішером. Від'ємний знак кореляції означає згоду: місце 1 у зовнішньому рейтингу відповідає найманіпулятивнішому джерелу категорії.

Похибку індексу джерела оцінено в такий спосіб: повідомлення кожного вікна випадково поділено на дві половини, індекс (4) обчислено на кожній, а модуль різниці усереднено за моментами часу. Семиденне вікно сталої ширини в середньому точніше за правило (3) (похибка 0,0197 проти 0,0304), бо містить більше повідомлень, проте похибка між джерелами різнилася в ньому вдвічі сильніше. Перевагу віддано зіставній між джерелами точності, а не найвищій середній: індекс визначає штраф для конкретного джерела в конкретний момент, і нерівна надійність оцінки означала б різний за строгістю штраф для різних джерел.

Метод відбору

Для запиту, що поєднує інформаційну потребу користувача u і момент t , з множини кандидатів $A(u, t)$ – повідомлень, доступних на цей момент, – обирається набір із k повідомлень з найбільшою сумою значень цільової функції (5):

$$S^* = \arg \max_{\substack{S \subseteq A(u, t) \\ |S|=k}} \sum_{a \in S} f(a), \quad (5)$$

де S^* – відібраний набір; $f(a)$ – значення цільової функції для повідомлення a .

У дослідженні цільова функція не містить складника залучення (цільової дії користувача), щоб виокремити взаємодію релевантності та штрафу за маніпулятивність. Значення $f(a)$ кожного кандидата не залежить від того, які ще кандидати увійшли до набору, тож задача (5) розв'язується точно: набір S^* складають k кандидатів із найбільшими значеннями $f(a)$. Без урахування джерела цільова функція поєднує релевантність і штраф за маніпулятивність повідомлення (6):

$$f_0(a) = \alpha_R R(u, a) - \alpha_M M(a), \quad (6)$$

де $R(u, a)$ – релевантність повідомлення a інформаційній потребі u ; $M(a)$ – показник маніпулятивності (1); $\alpha_R, \alpha_M \geq 0$ – вагові коефіцієнти релевантності та маніпулятивності.

Індекс джерела можна ввести до (6) двома способами: в адитивній формі (7) він входить окремим доданком, у мультиплікативній (8) – множить штраф за маніпулятивність повідомлення:

$$f_A(a) = \alpha_R R(u, a) - \alpha_M M(a) - \beta P_{src(a)}(t), \quad (7)$$

$$f_M(a) = \alpha_R R(u, a) - \left(1 + \beta P_{src(a)}(t)\right) \alpha_M M(a), \quad (8)$$

де $P_{src(a)}(t)$ – індекс (4) джерела повідомлення в момент запиту; $\beta \geq 0$ – вага індексу; решту позначень пояснено до (6).

Форми (7) і (8) різняться тим, як індекс діє на повідомлення різної маніпулятивності, що видно з похідних (9):

$$\frac{\partial f_A(a)}{\partial P_{src(a)}(t)} = -\beta, \quad \frac{\partial f_M(a)}{\partial P_{src(a)}(t)} = -\beta \alpha_M M(a), \quad \frac{\partial f_M(a)}{\partial M(a)} = -\left(1 + \beta P_{src(a)}(t)\right) \alpha_M, \quad (9)$$

де перші дві похідні – зміна цільової функції за зростання індексу джерела на одиницю, третя – граничний штраф за одиницю маніпулятивності в мультиплікативній формі.

В адитивній формі індекс понижує всі повідомлення насиченого джерела однаково, зокрема неманіпулятивні, як і оцінка якості джерела, розглянута в аналізі публікацій. У мультиплікативній формі пониження пропорційне маніпулятивності повідомлення: граничний штраф за одиницю маніпулятивності стає власним для кожного джерела, а повідомлення з $M(a) = 0$ індекс не понижує. Як ці властивості позначаються на складі показу, перевірено в експерименті.

МЕТОДИКА ДОСЛІДЖЕННЯ

Експеримент є позамережовим порівнянням конфігурацій функції відбору (5) на замороженому корпусі, а не оцінюванням користі для читача. Використано корпус експерименту (125 987 повідомлень 34 джерел за 182 доби) та індекс (4) за $n_{\min} = 10$. Перевірку побудовано за схемою тестових колекцій інформаційного пошуку: заморожений корпус, фіксований набір інформаційних потреб і оцінки релевантності, на яких конфігурації системи порівнюються між собою, а не за абсолютною якістю [16]. Інформаційну потребу u у (5) замінює тематичний профіль – нормований центроїд одного з $K = 12$ кластерів методу K -середніх на векторних поданнях повідомлень, обчислених моделлю multilingual-e5-large-instruct [17]. На відміну від класичних колекцій, теми не сформульовано людьми, а виведено



кластеризацією з самого корпусу. Запит поєднає профіль і момент t ; запитів $12 \times 182 = 2184$. Кандидати – повідомлення за 48 год до t із джерел із визначеним індексом, у середньому 1381 на запит; показується $k = 10$ повідомлень.

Релевантність – косинус між поданнями профілю й повідомлення, лінійно нормований за процентилями (10):

$$\tilde{R}(u, a) = \frac{\cos(\mathbf{e}(u), \mathbf{e}(a)) - \cos_{0,01}}{\cos_{0,99} - \cos_{0,01}}, \quad (10)$$

де $\tilde{R}(u, a)$ – нормована релевантність; $\mathbf{e}(u)$ – нормований центроїд профілю; $\mathbf{e}(a)$ – подання повідомлення; $\cos_{0,01}$, $\cos_{0,99}$ – 1-й і 99-й процентилі косинуса за всіма парами «профіль – кандидат».

Косинусна подібність між профілем і кандидатами без нормування може лежати у вузькому проміжку, тож відмінності релевантності між кандидатами були б малими; нормування (10) розтягує цей проміжок приблизно до $[0; 1]$ за єдиною для всіх запитів шкалою. У (6)–(8) $R(u, a) = \tilde{R}(u, a)$, $\alpha_R = 1$. Конфігурації відбору наведено в табл. 2; кожне порівняння ізолює один складник методу. У конфігурації Static замість поточного індексу використано сталий показник джерела (11):

$$\bar{P}_c = \frac{1}{|T_c|} \sum_{t \in T_c} P_c(t), \quad (11)$$

де $\bar{P}_c$ – сталий показник джерела c ; T_c – множина моментів, у яких індекс джерела c визначений; $P_c(t)$ – індекс (4).

Таблиця 2

Конфігурації відбору				
Конфігурація	Штраф у цільовій функції	Значення параметра	Порівнюється з	Що ізолює
Base	$-(\alpha_M = 0)$	–	–	точка відліку
Plain(α)	$\alpha_M M(a)$, (6)	$\alpha_M = 0,5; 1; 1,5; 2; 3; 5; 8$	Base	ціна штрафу без індексу; крива для порівняння за однакової ціни
Add(β)	$M(a) + \beta P_{\text{src}(a)}(t)$, (7)	$\beta = 0,5; 1; 2; 5; 10; 20$	Plain за однакової ціни; Mult	внесок адитивного врахування індексу
Mult(β)	$(1 + \beta P_{\text{src}(a)}(t)) M(a)$, (8)	ті самі β	Plain за однакової ціни; Add	внесок мультиплікативного врахування індексу
Static(β)	$(1 + \beta \bar{P}_{\text{src}(a)}) M(a)$, (11)	ті самі β	Mult(β)	внесок зміни індексу в часі
Source(β)	$\beta P_{\text{src}(a)}(t)$	$\beta = 0,05\text{--}20$	Plain, Add і Mult за однакової ціни	штраф лише джерела, без оцінки повідомлення

У табл. 2 α у назві Plain(α) – значення α_M ; в Add, Mult і Static $\alpha_M = 1$.

Склад показу описують дві частки (12):

$$\begin{aligned} \text{SatShare}_q(X) &= \frac{1}{k} \sum_{a \in S_q(X)} s_q(a), \\ \text{SatClean}_q(X) &= \frac{1}{k} \sum_{a \in S_q(X)} s_q(a) \mathbb{1}[M(a) \leq M_{0,25}], \\ s_q(a) &= \mathbb{1}[\text{src}(a) \in C_q^+], \end{aligned} \quad (12)$$

де $\text{SatShare}_q(X)$ – частка позицій насичених джерел; $\text{SatClean}_q(X)$ – частка позицій неманіпулятивних повідомлень насичених джерел; X – конфігурація; $S_q(X)$ – k повідомлень, відібраних нею для запиту q ; C_q^+ – насичені джерела запиту, тобто верхня чверть джерел кандидатів за $P_c(t)$; $s_q(a)$ – ознака того, що повідомлення походить із насиченого джерела; $M_{0,25}$ – перший квантиль M у корпусі, не вище якого повідомлення вважаються неманіпулятивними; $\mathbb{1}[\cdot]$ – індикатор.

Щоб охарактеризувати конфігурацію загалом, міри (12) усереднено за всіма запитами: $\text{SatShare}(X)$ і $\text{SatClean}(X)$ – середні значення $\text{SatShare}_q(X)$ і $\text{SatClean}_q(X)$. Так само за запитами усереднено $\text{Manip}@k$ – середнє $M(a)$ показаних повідомлень, $\text{Rel}@k$ – середнє $\tilde{R}(u, a)$ і $\text{Sources}@k$ – кількість різних джерел у показі. Міри (12) належать до поширених показників складу ранжованої видачі: частку груп серед перших



k позицій використовують для оцінювання справедливості ранжування [18], а зважений за позицією варіант цієї частки – для оцінювання нормативного різноманіття рекомендацій новин [8]. Обидві форми врахування індексу відрізняються від штрафу без індексу тим, що повідомлення насичених джерел понижуються сильніше, тож їхній внесок має виявитися в перерозподілі позицій між джерелами; SatClean показує, чи не витісняються при цьому неманіпулятивні повідомлення насичених джерел.

Ціну в релевантності задає (13):

$$\text{Cost}(X) = 100 \% \cdot \left(1 - \frac{\text{Rel}@k(X)}{\text{Rel}@k(\text{Base})} \right), \quad (13)$$

де $\text{Cost}(X)$ – на скільки відсотків показаний набір конфігурації X менш релевантний, ніж k найрелевантніших кандидатів (Base).

Посилення штрафу прибирає з показу дедалі більше повідомлень насичених джерел, тому кожну форму порівнюють зі штрафом без індексу тієї самої ціни за контрастом – різницею значень міри (14):

$$\begin{aligned} \Delta_F(\beta) &= \text{SatShare}(F(\beta)) - g(C_F(\beta)), \quad C_F(\beta) = \text{Cost}(F(\beta)), \\ \delta_F(\beta) &= 100 \% \cdot \frac{\Delta_F(\beta)}{g(C_F(\beta))}, \quad F \in \{\text{Add}, \text{Mult}\}, \end{aligned} \quad (14)$$

де $\Delta_F(\beta)$ – внесок урахування індексу у формі F за параметра β (далі Δ_{add} і Δ_{mult}); $\delta_F(\beta)$ – той самий внесок відносно рівня штрафу без індексу тієї самої ціни, % (далі δ_{add} і δ_{mult}); $C_F(\beta)$ – ціна конфігурації $F(\beta)$; $g(c)$ – частка позицій насичених джерел, якої штраф без індексу досягає за ціни c , отримана лінійною інтерполяцією між точками $(\text{Cost}(X), \text{SatShare}(X))$ конфігурацій Base і Plain(α) без екстраполяції.

Ті самі контрасти обчислено для SatClean і Manip@ k . Відносну форму δ_F використано тому, що штраф без індексу вже за ціни близько 1 % лишає насиченим джерелам лише близько 0,055 частки позицій, тож абсолютний контраст за SatShare за модулем не може перевищити цієї величини. Внесок зміни індексу в часі оцінено за однакового β контрастом (15):

$$\Delta_{\text{time}}(\beta) = \text{SatShare}(\text{Mult}(\beta)) - \text{SatShare}(\text{Static}(\beta)), \quad (15)$$

де $\Delta_{\text{time}}(\beta)$ – внесок того, що індекс змінюється в часі.

Для того самого запиту список $S = S_q(X)$ порівнюють з опорним $S' = S_q(Y)$ за мірами (16):

$$\begin{aligned} \text{Repl}_q(X, Y) &= |S \setminus S'|, \\ \text{SatShift}(X, Y) &= \frac{\sum_q \sum_{a \in G_q} (\pi_s(a) - \pi_{s'}(a))}{\sum_q |G_q|}, \\ G_q &= \{a \in S \cap S' : s_q(a) = 1\}, \end{aligned} \quad (16)$$

де Repl_q – кількість замінів, тобто повідомлень списку S , яких немає в опорному списку S' ; $\pi_s(a)$ – позиція повідомлення a у списку S , упорядкованому за спаданням $f(a)$; G_q – повідомлення насичених джерел, що лишилися в обох списках; SatShift – їхній середній зсув, додатний – униз.

Для витіснених повідомлень (ϵ в S' , немає в S) і нових (навпаки) обчислено середню позицію, частку з насичених джерел і середнє M .

Одиниця аналізу – запит; 95 % довірчі інтервали (ДІ) отримано циклічним блоковим бутстрепом за днями [19] з блоками по 7 діб і 10 000 повторами, бо 48-годинні вікна кандидатів роблять сусідні дні залежними. Ефект вважається статистично значущим на рівні 0,05, якщо його 95 % ДІ не містить нуля. Контрасти (14) повторено у 12 варіантах умов: змінено кількість профілів (8 і 20), початковий стан K -середніх (два варіанти), вікно кандидатів (24 год), k (5 і 20), модель векторного подання (ukr-paraphrase-multilingual-mpnet-base [20] для всього тексту й для першого фрагмента, Qwen3-Embedding-0.6B [21]) і нормування косинуса (два способи).

Окремо перевірено залежність результатів від β і їхню стійкість до повторів однієї новини: та сама подія часто потрапляє до показу кількома повідомленнями різних джерел. Для другої перевірки відбір повторено так, щоб кожну новину представляло одне повідомлення: кандидати переглядаються в порядку спадання $f(a)$, і кандидат береться, якщо не повторює жодного вже взятого. Пара повідомлень a , b вважається повтором, якщо їхні тексти збігаються дослівно або виконується умова (17):

$$\cos(\mathbf{v}(a), \mathbf{v}(b)) \geq 0,78 \quad \text{і} \quad \frac{|W(a) \cap W(b)|}{\min(|W(a)|, |W(b)|)} \geq 0,30, \quad (17)$$

де $\mathbf{v}(a)$ – подання повідомлення a моделлю ukr-paraphrase-multilingual-mpnet-base, донавченою для української мови [20] версією багатомовної парафразної моделі [22]; $W(a)$ – множина змістових слів і чисел повідомлення a .

На відміну від вилучення дослівних повторів під час перевірки індексу (табл. 1), умова (17) виділяє



повідомлення про ту саму подію чи заяву незалежно від стилю; службові повідомлення про повітряні тривоги й рух засобів ураження, які різняться лише місцем і часом, повторами не вважаються. На контрольній вибірці зі 150 пар, розміченій мовною моделлю, точність правила (17) становить 0,873 (95 % ДІ [0,76; 0,96]), повнота – 0,66 [0,50; 0,83].

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Табл. 3 зіставляє конфігурації однакової ціни без інтерполяції: ціну Plain(1,5), 1,09 %, адитивна форма досягає за $\beta = 0,5$, мультиплікативна – за $\beta = 10$, штраф лише джерела – за $\beta = 1,46$. Поряд наведено обидві форми з $\beta = 5$. Для Base наведено абсолютні значення мір, для решти конфігурацій – їхню зміну відносно Base; Cost і Repl(X, Base) відрховуються від Base за означенням.

Таблиця 3

Ціна штрафу і склад показу, $n_{\min} = 10, k = 10$

Конфігурація	Cost, %	Manip@10	SatShare	SatClean	Sources@10	Repl(X, Base)
Base	0	0,0423	0,085	0,009	6,70	0
<i>однакова ціна</i>						
Plain(1,5)	1,09	–81 %	–35 %	+87 %	–2 %	3,51
Add(0,5)	1,08	–78 %	–76 %	–39 %	–9 %	3,69
Mult(10)	1,09	–81 %	–45 %	+84 %	–3 %	3,48
Static(10)	1,09	–81 %	–42 %	+85 %	–4 %	3,48
Source(1,46)	1,09	–32 %	–93 %	–94 %	–19 %	3,82
$\beta = 5$						
Add(5)	4,68	–81 %	–100 %	–100 %	–36 %	6,88
Mult(5)	0,97	–79 %	–40 %	+80 %	–3 %	3,28

Більшу частину ефекту дає штраф без індексу: Plain(1,5) знижує Manip@10 на 81 %, тобто в 5,4 раза, за ціни 1,09 %. Ціна мала, бо під десятикою найрелевантніших кандидатів лежить багато майже так само релевантних. За однакової ціни Plain(1,5), Static(10) і Mult(10) однаково знижують маніпулятивність показу, а різняться часткою позицій насичених джерел. Неманіпулятивним повідомленням насичених джерел у цих конфігураціях дістається майже вдвічі більше позицій, ніж у Base (+84–87 %). Адитивна форма за тієї самої ціни зменшує частку позицій насичених джерел сильніше, ніж мультиплікативна (–76 % проти –45 %), але скорочує й частку їхніх неманіпулятивних повідомлень (–39 %), тоді як мультиплікативна її збільшує (+84 %).

Таблиця 4

Відносні контрасти урахування індексу $\delta_F(\beta)$ (14), % [95 % ДІ], $n_{\min} = 10, k = 10$

Контраст	Cost, %	SatShare	SatClean	Manip@10
$\delta_{\text{mult}}, \beta = 5$	0,97	–9,8 [–11,9; –7,9]	–0,1 [–1,5; 1,1]	–0,6 [–0,7; –0,5]
$\delta_{\text{mult}}, \beta = 10$	1,09	–15,6 [–19,3; –12,6]	–1,4 [–3,1; –0,1]	+1,3 [1,0; 1,5]
$\delta_{\text{add}}, \beta = 0,5$	1,08	–62,8 [–70,1; –56,0]	–67,5 [–78,5; –57,5]	+17,5 [16,5; 18,4]
$\delta_{\text{add}}, \beta = 1$	1,48	–83,9 [–89,9; –77,5]	–86,5 [–93,5; –78,8]	+52,1 [49,8; 54,5]
$\delta_{\text{add}}, \beta = 2$	2,36	–94,9 [–98,5; –90,6]	–96,8 [–99,6; –93,1]	+136,6 [129,8; 143,8]

Відносно штрафу без індексу тієї самої ціни мультиплікативна форма за $\beta = 5$ зменшує частку позицій насичених джерел на 9,8 %, не змінюючи частки їхніх неманіпулятивних повідомлень (табл. 4), а за $\beta = 10$, коли її ціна така сама, як в Add(0,5), – на 15,6 %, витісняючи 1,4 % їхніх неманіпулятивних повідомлень. Адитивна форма за $\beta = 0,5$ зменшує частку позицій насичених джерел на 62,8 %, але на 67,5 % скорочує й частку їхніх неманіпулятивних повідомлень, а Manip@10 у неї на 17,5 % вищий: частину ціни в релевантності вона витрачає на пониження джерел, а не маніпулятивних повідомлень. З ростом β контраст мультиплікативної форми за модулем зростає (рис. 1, а) і за $\beta = 20$ сягає –24,5 %; за $n_{\min} = 30$ абсолютний контраст Δ_{mult} за $\beta = 5$ становить –0,0069 [–0,0079; –0,0059] проти –0,0055 [–0,0064; –0,0047]. В усіх 12 варіантах умов інтервали контрастів за SatShare не містять нуля: для мультиплікативної форми за $\beta = 5$ оцінки лежать від –8,5 до –24,8 %, для адитивної за $\beta = 0,5$ – від –50 до –100 %. За $\beta \geq 5$ ціна адитивної форми (4,7–10 %) перевищує ціну найсильнішого штрафу без індексу Plain(8), 2,8 %, тож контраст для неї не визначений. Часова складова Δ_{time} за $\beta = 5$, –0,0012 [–0,0017; –0,0007], дорівнює близько п'ятої частини контрасту Δ_{mult} , тобто мультиплікативна форма діє переважно через сталу властивість джерела, а не через зміну індексу в часі.

Від β залежить компроміс між зсувом показу від насичених джерел і збереженням їхніх неманіпулятивних повідомлень (рис. 1): до $\beta = 5$ мультиплікативна форма за тієї самої ціни не витісняє неманіпулятивних повідомлень насичених джерел понад штраф без індексу, а Manip@10 навіть трохи нижчий, ніж у Plain. З $\beta = 10$ з'являється мале, але статистично значуще додаткове витіснення, яке далі зростає, а Manip@10 стає трохи вищим, ніж у Plain. Ця межа залежить від k : за $k = 5$ витіснення помітне вже за $\beta = 5$ (рис. 1, б). Для адитивної форми контраст визначений лише за β від 0,5 до 2 (табл. 4): з ростом β вона дедалі повніше витісняє насичені джерела разом з їхніми неманіпулятивними повідомленнями, а Manip@10 зростає.

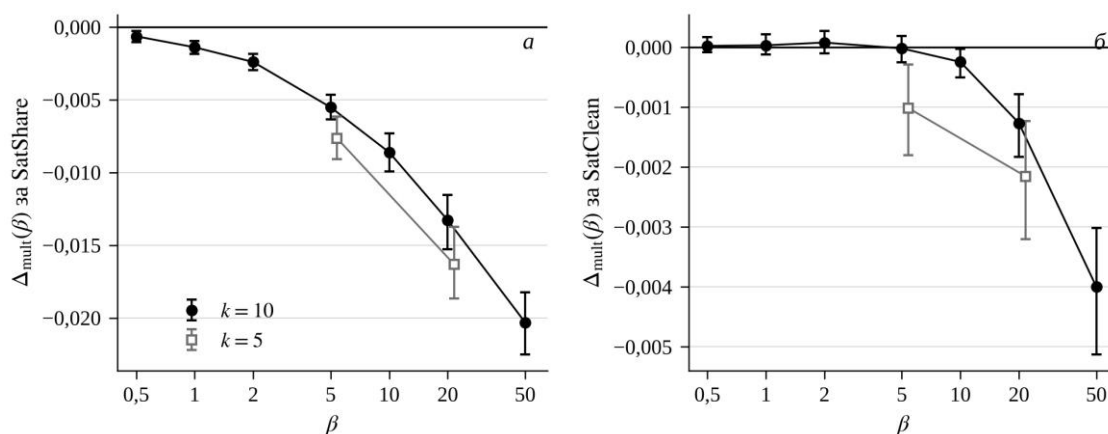


Рис. 1. Контраст мультиплікативної форми $\Delta_{mult}(\beta)$ з 95 % ДІ залежно від β для $k = 10$ і $k = 5$: а – за SatShare; б – за SatClean

Штраф без індексу замінює в середньому 3,5 повідомлення з 10 (табл. 5), і витіснені повідомлення значно маніпулятивніші за нові. Мультиплікативна форма змінює показ здебільшого перестановкою, а заміни припадають на 9–10-ту позиції. Серед витіснених частка повідомлень насичених джерел більша, ніж серед нових, проте нові не менш маніпулятивні. Отже, за однакової ціни мультиплікативна форма не зменшує маніпулятивності показу, а розподіляє штраф між джерелами: з двох однаково маніпулятивних повідомлень витісняється те, що походить із насиченого джерела, а повідомлення насичених джерел, що лишилися, опускаються в середньому на 0,72 позиції. Замін між Mult(10) і Static(10) удвічі менше, ніж між Mult(10) і Plain(1,5).

Таблиця 5

Зміни показу відносно Base і між конфігураціями однакової ціни, $n_{min} = 10$

Показник	Plain(1,5) проти Base	Mult(10) проти Plain(1,5)	Mult(10) проти Static(10)	Add(0,5) проти Mult(10)
Що ізолює порівняння	штраф	мультиплікативну форму	зміну індексу в часі	форму врахування індексу
Списки із замінами, %	95,3	35,1	18,6	82,0
Замін на список, Repl	3,51	0,38	0,19	1,39
Середня позиція: витіснені / нові	6,0 / 7,5	9,4 / 9,4	9,6 / 9,6	8,0 / 8,6
З насичених джерел, %: витіснені / нові	14,3 / 5,9	23,2 / 0,4	15,7 / 4,1	19,6 / 0,8
Середнє M: витіснені / нові	0,104 / 0,006	0,015 / 0,019	0,016 / 0,019	0,006 / 0,015
SatShift, позицій	-0,70	+0,72	+0,23	+2,28

Порівняно з Mult(10) адитивна форма Add(0,5) замінює 1,39 повідомлення на список, і заміни припадають на вищі позиції: витіснені повідомлення мають $M \approx 0,006$, а нові маніпулятивніші ($M \approx 0,015$) і майже не походять із насичених джерел (табл. 5). За однакового β від 0,5 до 20 частка неманіпулятивних повідомлень насичених джерел в адитивній формі нижча, ніж у мультиплікативній, а Add(5) майже прибирає насичені джерела з показу і зменшує кількість різних джерел у показі на 36 % проти 3 % у Mult(5); ціна адитивної форми зростає з β і за $\beta = 20$ сягає 10 %. Крайній випадок – штраф



лише джерела, $Source(\beta)$: за тієї самої ціни (1,09 %) $Source(1,46)$ майже прибирає насичені джерела разом з їхніми неманіпулятивними повідомленнями (–93 % і –94 %), найсильніше скорочує кількість різних джерел у показі (–19 %) і замінює найбільше повідомлень (3,82), а $Manip@10$ знижує лише на 32 %. Перевага мультиплікативної форми над штрафом лише джерела стійка: за всіх β від 1 до 20 $Mult(\beta)$ зберігає більше неманіпулятивних повідомлень насичених джерел і сильніше знижує $Manip@10$, ніж $Source$ тієї самої ціни, і обидві різниці статистично значущі.

Оскільки ціну (13) вимірює та сама модель векторного подання, що визначає релевантність у цільовій функції, її перевірено незалежними оцінками великої мовної моделі. Такі оцінки відтворюють упорядкування систем за ручними оцінками [23], хоча відмінності між системами розрізняють гірше [24]. Оцінювач релевантності – Claude Sonnet 5.5 [25], не пов'язаний ні з моделлю векторного подання, ні з класифікатором, – ставить парі «профіль – кандидат» оцінку тематичної доречності від 0 до 3; профіль подано 10 найближчими до центроїда повідомленнями, кандидата – текстом без позначок джерела. Для 120 запитів оцінено 2 259 таких пар і ще 109 пар з показу $Add(0,5)$, яких серед них не було, а ціну за оцінками обчислено за (18):

$$Cost_{nDCG}(X) = 100 \% \cdot \left(1 - \frac{nDCG@10(X)}{nDCG@10(Base)} \right), \quad (18)$$

де $nDCG@10(X)$ – нормований дисконтований кумулятивний виграш [26] перших 10 позицій конфігурації X ; виграш повідомлення – середнє двох оцінок, отриманих за двома формулюваннями інструкції.

Оцінки узгоджуються з релевантністю (10): вона розрізняє пари з оцінками 2–3 і 0–1 з площею під ROC-кривою (AUC) 0,907. Від формулювання інструкції вони майже не залежать: коефіцієнт каппа з квадратичними вагами, що дорівнює 1 за повного збігу й 0 за збігу на рівні випадкового, становить 0,877; зміщення оцінок за маніпулятивністю не виявлено. За оцінками ціна приблизно вдвічі більша, ніж за (13), але впорядкування конфігурацій за ціною те саме (табл. 6).

Таблиця 6

Ціна штрафу, виміряна за релевантністю (10) і за незалежними оцінками, $k = 10$

Конфігурація	Cost (13), %	nDCG@10 [95 % ДІ]	Cost _{nDCG} (18), % [95 % ДІ]
Base	0	0,908 [0,893; 0,922]	0
Plain(1)	0,84	0,893 [0,874; 0,912]	1,64 [0,54; 2,83]
Mult(1)	0,87	0,891 [0,871; 0,911]	1,82 [0,70; 3,06]
Mult(5)	0,97	0,890 [0,870; 0,910]	1,99 [0,80; 3,32]
Add(0,5)	1,08	0,890 [0,870; 0,909]	2,01 [0,81; 3,31]
Mult(20)	1,28	0,883 [0,861; 0,904]	2,72 [1,34; 4,17]

Без вилучення повторів одна новина часто займає в показі кілька позицій: у показі Base у середньому 7,35 різних новин. Якщо кожну новину представляє одне повідомлення (17), результати зберігаються: відносний контраст мультиплікативної форми за $\beta = 5$ становить –9,6 % [–11,7; –7,8] проти –9,8 %, і за жодної зі змін умов інтервал не містить нуля; для адитивної форми за $\beta = 0,5$ він становить –66,5 % [–71,3; –61,5] проти –62,8 %, і його інтервал також не містить нуля за жодної зі змін умов.

Обговорення та обмеження

Більшу частину зниження маніпулятивності показу дає штраф за маніпулятивність конкретного повідомлення самотужки: за ціни 1,09 % він знижує $Manip@10$ на 81 % і частку позицій насичених джерел на 35 %, тож обидві форми врахування індексу діють на залишок, що лишається після нього. Штраф лише джерела (конфігурація $Source$) на цих даних поступається штрафу за маніпулятивність повідомлення і за зниженням $Manip@10$, і за збереженням неманіпулятивних повідомлень насичених джерел. Обидві форми врахування індексу переважно перерозподіляють позиції між джерелами, але по-різному. Мультиплікативна понижує повідомлення пропорційно до їхньої маніпулятивності: з двох однаково маніпулятивних повідомлень вище опиняється те, що походить із менш насиченого джерела, а неманіпулятивні повідомлення насичених джерел зберігають позиції. Адитивна понижує всі повідомлення насиченого джерела однаково, тож сильніше зміщує показ від таких джерел, але витісняє й їхні неманіпулятивні повідомлення, а маніпулятивність показу знижує менше, ніж штраф без індексу тієї самої ціни. Вибір форми залежить від того, що важливіше в конкретній системі: сильніше зменшити присутність насичених джерел чи зберегти доступ до їхніх неманіпулятивних повідомлень. Внесок зміни індексу в часі малий, тому сталий показник джерела $\bar{P}_c$ дав би майже той самий перерозподіл; проте $\bar{P}_c$ усереднює індекс за весь період, зокрема за моменти після запиту, і під час роботи системи недоступний. Індекс (4) спирається лише на повідомлення, опубліковані до моменту запиту, тож дає практично той самий ефект без звернення до майбутніх даних. Наявність і напрям ефектів обох форм від умов експерименту не



залежать: в усіх 12 варіантах умов інтервали контрастів за SatShare не містять нуля.

Результати мають три обмеження. По-перше, це позамережеве порівняння конфігурацій на замороженому корпусі з темами, виведеними з самого корпусу: воно показує, як змінюється склад показу, але не вимірює впливу на читача. По-друге, показник $M(a)$ у мірах складу показу обчислює той самий класифікатор, що й у штрафі, тому зниження $\text{Map}@10$ закладене побудовою, а якість класифікатора виміряно на випробувальній вибірці UNLP 2025, тоді як на корпусі 2026 року, зібраному з інших джерел, вона не вимірювалася. По-третє, перевірка узгодженості індексу із зовнішнім рейтингом не є цілком незалежною: набір даних, на якому навчено класифікатор, підготувала та сама редакція [5], тож незалежними є інструмент, період і виконавці вимірювання, але не поняття маніпулятивності.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Адитивне й мультиплікативне врахування маніпулятивної насиченості джерела по-різному розподіляють штраф між повідомленнями насиченого джерела: мультиплікативна форма робить граничний штраф за одиницю маніпулятивності власним для кожного джерела й не понижує неманіпулятивних повідомлень, адитивна понижує всі повідомлення джерела однаково. На корпусі 125 987 повідомлень 34 джерел за 182 доби маніпулятивність показу знижує передусім штраф за маніпулятивність самого повідомлення – у 5,4 раза ціною 1,09 % релевантності (за незалежними оцінками ціна приблизно вдвічі більша). За тієї самої ціни адитивна форма сильніше зменшує частку позицій насичених джерел (–76 % проти –45 % відносно відбору лише за релевантністю), але скорочує й частку їхніх неманіпулятивних повідомлень (–39 %), тоді як мультиплікативна її збільшує (+84 %). Відносно штрафу без індексу тієї самої ціни (1,09 %) мультиплікативна форма зменшує частку позицій насичених джерел на 15,6 %, адитивна – на 62,8 %, але адитивна при цьому менше знижує маніпулятивність показу; обидва ефекти статистично значущі й зберігаються у 12 варіантах умов і за видалення повторів однієї новини.

Індекс, визначений як середня маніпулятивність повідомлень джерела у вікні, що розширюється до заданої кількості повідомлень, придатний як вимірювальний засіб: він розрізняє джерела (мінливість між джерелами в 3,48 раза перевищує мінливість у часі), стійкий до параметрів обчислення (рангова кореляція не нижча за 0,965), відтворюється на незалежних половинах даних (0,9962), узгоджується з незалежним вимірюванням і має вдвічі менший розкид точності між джерелами, ніж вікна сталої ширини. Він обчислюється під час обробки запиту лише за минулими повідомленнями джерела й не потребує даних про користувачів.

Перспективами подальших досліджень є перевірка впливу відбору на читачів в умовах реального показу та заміна маніпулятивного повідомлення його переписаним варіантом із меншою маніпулятивністю з вимірюванням того, як така заміна змінює маніпулятивну насиченість джерела.

ДЕКЛАРАЦІЯ ПРО ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ

Під час підготовки рукопису автори використовували Claude Opus 5.5 (Anthropic) для редагування тексту й перевірки його на помилки. Усі запропоновані зміни автори переглянули й перевірили та несуть повну відповідальність за зміст статті.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
2. Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3), pgaf062. <https://doi.org/10.1093/pnasnexus/pgaf062>
3. European Parliament and Council of the European Union. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
4. Кельм, Н., Романишин, Н., Гебура, О., Міхальков, С., & Дукач, Ю. (2024). *Карусель емоцій. Рівень маніпулятивності мови українських телеграм-каналів*. Texty.org.ua. <https://texty.org.ua/projects/113607/karusel-emocij-riven-manipulyatyvnosti-movy-ukrayinskyh-telegram-kanaliv/>



5. Kyslyi, R., Romanyshyn, N., & Sydorskyi, V. (2025). The UNLP 2025 shared task on detecting social media manipulation. *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, 105–111. <https://doi.org/10.18653/v1/2025.unlp-1.12>
6. Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, 7(8), 993–1012. <https://doi.org/10.1080/21670811.2019.1623700>
7. Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., & Helberger, N. (2021). Recommenders with a mission: Assessing diversity in news recommendations. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, 173–183. <https://doi.org/10.1145/3406522.3446019>
8. Vrijenhoek, S., Bénédicte, G., Gutierrez Granada, M., Odijk, D., & de Rijke, M. (2022). RADio: Rank-aware divergence metrics to measure normative diversity in news recommendations. *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys 2022)*, 208–219. <https://doi.org/10.1145/3523227.3546780>
9. Jia, C., Lam, M. S., Mai, M. C., Hancock, J. T., & Bernstein, M. S. (2024). Embedding democratic values into social media AIs via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 163:1–163:36. <https://doi.org/10.1145/3641002>
10. Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 390(6776), eadu5584. <https://doi.org/10.1126/science.adu5584>
11. Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7), 2521–2526. <https://doi.org/10.1073/pnas.1806781116>
12. Loftsgaarden, D. O., & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3), 1049–1051. <https://doi.org/10.1214/aoms/1177700079>
13. INSAIT. (2025). *MamayLM-Gemma-3-4B-IT-v1.0*. Hugging Face model repository. <https://huggingface.co/INSAIT-Institute/MamayLM-Gemma-3-4B-IT-v1.0>
14. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffmann, A., & Giovannini, E. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
15. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
16. Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4), 247–375. <https://doi.org/10.1561/1500000009>
17. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. *arXiv*. <https://doi.org/10.48550/arXiv.2402.05672>
18. Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*, 2221–2231. <https://doi.org/10.1145/3292500.3330691>
19. Politis, D. N., & Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In R. LePage & L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 263–270). Wiley.
20. Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP 2023)*, 11–19. <https://doi.org/10.18653/v1/2023.unlp-1.2>
21. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv*. <https://doi.org/10.48550/arXiv.2506.05176>
22. Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
23. Thomas, P., Spielman, S., Craswell, N., & Mitra, B. (2024). Large language models can accurately predict searcher preferences. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, 1930–1940. <https://doi.org/10.1145/3626772.3657707>
24. Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., & Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2023)*, 39–50. <https://doi.org/10.1145/3578337.3605136>



25. Anthropic. (2026). *System Card: Claude Sonnet 5.5*. <https://www.anthropic.com/claude-sonnet-5-5-system-card>
26. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446. <https://doi.org/10.1145/582415.582418>

**Vladyslav V. Hromenko**

PhD student

Institute of Software Systems of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

ORCID: 0009-0001-5285-9912

gromvlad12@gmail.com

Viktor L. Shevchenko

Doctor of Technical Sciences, Professor, Deputy Director for Research

Institute of Software Systems of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

Professor of the Department of Information and

Cyber Security named after Professor Volodymyr Buriachok

Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine

ORCID: 0000-0002-9457-7454

gii2014@ukr.net

**USING A SOURCE-LEVEL MANIPULATIVE SATURATION INDEX IN MULTI-CRITERIA
SELECTION OF NEWS MESSAGES**

Abstract. News selection systems that rank messages by engagement tend to amplify manipulative content, and manipulateness may be a stable characteristic of the editorial practice of particular sources. Existing approaches typically treat these factors separately: they either penalize a manipulative message equally regardless of its source, or demote a source as a whole together with its non-manipulative messages. The article aims to investigate additive and multiplicative ways of accounting for the manipulative saturation of the source in multi-criteria selection of news messages and to compare them on real data. Message manipulateness is estimated by an LLM-based multi-label classifier of ten manipulation techniques. The manipulative saturation index of a source is defined as the mean manipulateness of the source's messages in a time window and is validated as a measurement instrument. In the additive form, the index enters the objective function as a separate term and demotes all messages of the source; in the multiplicative form, it multiplies the penalty for message manipulateness and does not demote non-manipulative messages. Both forms are compared with the penalty without the index and with each other at equal relevance cost on a frozen corpus of 125,987 messages from 34 sources over 182 days for 2,184 topical queries. The penalty on the manipulateness of the message itself reduces the mean manipulateness of shown messages 5.4-fold at a cost of 1.09 % of relevance. At equal cost, the additive form shifts positions away from saturated sources more strongly but also displaces their non-manipulative messages; the multiplicative form barely affects non-manipulative messages but reduces the share of positions held by saturated sources more moderately: by 15.6 % versus 62.8 % for the additive form, relative to the penalty without the index at the same cost.

Keywords: machine learning; information environment; news messages; manipulation techniques; manipulation detection; large language models; recommender systems; ranking; multi-criteria optimization.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
2. Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3), pgaf062. <https://doi.org/10.1093/pnasnexus/pgaf062>
3. European Parliament and Council of the European Union. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
4. Kelm, N., Romanyshyn, N., Hebur, O., Mikhalkov, S., & Dukach, Y. (2024). *Karusel emotsii. Riven manipulyatynosti movy ukrainskykh telehram-kanaliv* [Carousel of emotions: The level of manipulateness of the language of Ukrainian Telegram channels]. Texty.org.ua. <https://texty.org.ua/projects/113607/karusel-emocij-riven-manipulyatynosti-movy-ukrayinskyh->



- [telehram-kanaliv/](#)
5. Kyslyi, R., Romanyshyn, N., & Sydorskyi, V. (2025). The UNLP 2025 shared task on detecting social media manipulation. *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, 105–111. <https://doi.org/10.18653/v1/2025.unlp-1.12>
 6. Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism*, 7(8), 993–1012. <https://doi.org/10.1080/21670811.2019.1623700>
 7. Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., & Helberger, N. (2021). Recommenders with a mission: Assessing diversity in news recommendations. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, 173–183. <https://doi.org/10.1145/3406522.3446019>
 8. Vrijenhoek, S., Bénédicte, G., Gutierrez Granada, M., Odijk, D., & de Rijke, M. (2022). RADio: Rank-aware divergence metrics to measure normative diversity in news recommendations. *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys 2022)*, 208–219. <https://doi.org/10.1145/3523227.3546780>
 9. Jia, C., Lam, M. S., Mai, M. C., Hancock, J. T., & Bernstein, M. S. (2024). Embedding democratic values into social media AIs via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 163:1–163:36. <https://doi.org/10.1145/3641002>
 10. Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 390(6776), eadu5584. <https://doi.org/10.1126/science.adu5584>
 11. Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7), 2521–2526. <https://doi.org/10.1073/pnas.1806781116>
 12. Loftsgaarden, D. O., & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3), 1049–1051. <https://doi.org/10.1214/aoms/1177700079>
 13. INSAIT. (2025). *MamayLM-Gemma-3-4B-IT-v1.0*. Hugging Face model repository. <https://huggingface.co/INSAIT-Institute/MamayLM-Gemma-3-4B-IT-v1.0>
 14. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffmann, A., & Giovannini, E. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
 15. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
 16. Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4), 247–375. <https://doi.org/10.1561/1500000009>
 17. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. *arXiv*. <https://doi.org/10.48550/arXiv.2402.05672>
 18. Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*, 2221–2231. <https://doi.org/10.1145/3292500.3330691>
 19. Politis, D. N., & Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In R. LePage & L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 263–270). Wiley.
 20. Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP 2023)*, 11–19. <https://doi.org/10.18653/v1/2023.unlp-1.2>
 21. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv*. <https://doi.org/10.48550/arXiv.2506.05176>
 22. Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
 23. Thomas, P., Spielman, S., Craswell, N., & Mitra, B. (2024). Large language models can accurately predict searcher preferences. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, 1930–1940. <https://doi.org/10.1145/3626772.3657707>
 24. Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., & Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information*



- Retrieval (ICTIR 2023), 39–50. <https://doi.org/10.1145/3578337.3605136>
25. Anthropic. (2026). *System Card: Claude Sonnet 5.5*. <https://www.anthropic.com/claude-sonnet-5-5-system-card>
26. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446. <https://doi.org/10.1145/582415.582418>

Отримано редакцією журналу / Received: 21.08.26

Прорецензовано / Revised: 30.08.26

Схвалено до друку / Accepted: 26.10.26



This work is licensed under Creative Commons Attribution-noncommercial-sharealike 4.0 International License.