КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.33.1166
УДК 004.056

Дженджеро Дмитро Сергійович
Аспірант кафедри кібербезпеки та захисту інформації
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID: 0009-0007-9999-850X
dzhendzherod@gmail.com

Наконечний Володимир Сергійович
доктор технічних наук, професор, професор кафедри кібербезпеки та захисту інформації
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID: 0000-0002-0247-5400
volodym.nakonechnyi@knu.ua

МЕТОД АДАПТИВНОГО ВІДБОРУ ТА ЗВАЖУВАННЯ ОЗНАК НЕПРАВДИВОЇ ІНФОРМАЦІЇ ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ЇЇ ВИЯВЛЕННЯ В УМОВАХ ГІБРИДНОЇ ВІЙНИ

Анотація. У статті розв’язано актуальне для умов гібридної війни завдання підвищення ефективності виявлення неправдивої інформації в текстових повідомленнях. Актуальність дослідження зумовлена тим, що дезінформаційні впливи в сучасному інформаційному середовищі використовуються як інструмент підриву довіри до державних інституцій, викривлення сприйняття подій, дестабілізації суспільних настроїв і створення додаткового навантаження на системи прийняття рішень. У вступі обґрунтовано потребу в інтерпретованому та ресурсно-ефективному методі, придатному для роботи з великими, динамічними та незбалансованими потоками повідомлень. У постановці проблеми показано, що використання повного ознакового простору підвищує обчислювальну складність, ускладнює пояснення рішень і не враховує потребу в пріоритезації перевірки повідомлень за рівнем ризику. У розділі аналізу останніх досліджень і публікацій узагальнено сучасні підходи до виявлення неправдивої інформації, зокрема контентні, фактологічні, поведінкові та гібридні моделі, а також підходи до оцінювання дезінформаційних ризиків в умовах війни. У теоретичних основах дослідження систематизовано положення щодо відбору ознак, побудови ознакового простору, термового зважування, оцінювання якості класифікації та використання PR/ROC-подань в умовах дисбалансу класів. На цій основі сформовано концептуальну рамку методу адаптивного відбору та зважування ознак неправдивої інформації. У методиці дослідження описано експериментальну перевірку на відкритому українськомовному корпусі новин, тематично пов’язаних із подіями війни Російської Федерації проти України. Після очищення корпусу, вилучення порожніх записів і дублікатів, а також застосування фільтра довжини не менше 200 символів сформовано вибірку обсягом 29372 повідомлення, з яких 353 належать до класу неправдивої інформації, а 29019 – до класу правдивої інформації. Для побудови ознакового опису використано TF-IDF для уніграм і біграм, а як базову модель – логістичну регресію. Відбір ознак реалізовано за схемою χ^2 + top-K з перевіркою кількох значень K на валідаційній вибірці; робочим варіантом обрано K=5000. Для практичного використання скорингу введено трирівневий тріаж ризику на основі 80-го та 95-го процентилів скорингу. У розділі результатів показано, що скорочення ознакового простору з 30000 до 5000 ознак не призводить до суттєвого погіршення якості класифікації: на тестовій вибірці значення F1 зменшується з 0,768 до 0,760, а PR-AUC – з 0,819 до 0,805. Водночас тріаж ризику підтвердив практичну придатність методу: високоризикова група охопила 256 повідомлень із 4406 на тестовій вибірці та містила 50 із 53 фейків, тоді як у низькоризиковій групі зафіксовано лише 1 фейк. У висновках обґрунтовано, що запропонований метод може застосовуватися як інтерпретований та ресурсно-ефективний компонент систем моніторингу інформаційного простору, а подальші дослідження доцільно спрямувати на розширення набору ознак і перевірку методу на інших українськомовних корпусах.

Ключові слова: неправдива інформація; виявлення фейків; гібридна війна; відбір ознак; TF-IDF; логістична регресія; тріаж ризику; інформаційна безпека.

ВСТУП

Стрімка цифровізація інформаційного середовища, зростання ролі соціальних платформ і мережевих каналів комунікації, а також високі темпи поширення повідомлень істотно ускладнили завдання виявлення неправдивої інформації. В умовах гібридної війни ця проблема набуває не лише інформаційного, а й безпекового значення, оскільки дезінформаційні впливи використовуються для підриву довіри до державних інституцій, викривлення сприйняття подій, дестабілізації суспільних настроїв і створення додаткового навантаження на системи прийняття рішень. Для України така постановка є особливо актуальною, оскільки інформаційне протиборство супроводжує воєнні дії, політичні процеси, гуманітарні питання та функціонування критично важливих цифрових сервісів.

Сучасні дослідження у сфері виявлення неправдивої інформації охоплюють широкий спектр підходів: від аналізу змісту тексту, стилістичних і лінгвістичних маркерів до оцінювання джерела, фактологічного зіставлення, поведінкових характеристик поширення та побудови гібридних моделей, які поєднують кілька груп сигналів. Разом з тим застосування таких підходів у прикладних умовах супроводжується низкою обмежень. По-перше, значна частина методів орієнтована на доступ до широкого набору платформних даних, що не завжди можливо в реальному середовищі. По-друге, складні моделі часто мають нижчий рівень інтерпретованості, що ускладнює пояснення рішень та їх використання у ризик-орієнтованих контурах реагування. По-третє, в умовах гібридної війни тематики, наративи та способи подання неправдивої інформації змінюються динамічно, а тому статичний набір ознак або жорстко фіксовані правила виявлення не завжди забезпечують достатню стійкість.

З огляду на це виникає потреба у методі, який поєднує кілька важливих властивостей: інтерпретованість, можливість адаптивного скорочення ознакового простору, збереження прийнятної якості виявлення та придатність до практичного використання в умовах обмежених ресурсів перевірки. Такий підхід має не лише класифікувати повідомлення, а й підтримувати пріоритезацію аналізу, тобто відокремлювати найбільш ризикові повідомлення для першочергової перевірки. У цьому контексті доцільним є поєднання відбору ознак, їх зважування та порогового тріажу ризику в межах єдиної процедури прийняття рішення.

У статті запропоновано метод адаптивного відбору та зважування ознак неправдивої інформації для підвищення ефективності її виявлення в умовах гібридної війни. Метод базується на використанні контентного ознакового опису текстів, евристичного скорочення ознакового простору, побудови скорингу належності до класу неправдивої інформації та подальшої трирівневої категоризації ризику. Експериментальну перевірку здійснено на відкритому українськомовному корпусі новин, тематично пов’язаних із подіями війни Російської Федерації проти України. Отримані результати дають змогу оцінити як вплив скорочення ознакового простору на метрики якості, так і практичну придатність тріажу для концентрації більшості фейків у невеликій частині інформаційного потоку.

Постановка проблеми. Проблема полягає у тому, що в умовах гібридної війни системи виявлення неправдивої інформації мають працювати з великими, динамічними та незбалансованими потоками текстових повідомлень, у яких кількість потенційно небезпечних повідомлень є відносно малою, але їх своєчасне виявлення має підвищену практичну цінність. Використання повного ознакового простору у таких умовах збільшує обчислювальну складність, ускладнює інтерпретацію рішень і може призводити до надлишковості ознак без суттєвого приросту якості. Водночас двокласове рішення без додаткової пріоритезації не враховує обмежені ресурси ручної перевірки та потребу в оперативному реагуванні. Отже, актуальним є розроблення методу, який забезпечує адаптивний відбір інформативних ознак, збереження прийнятної якості виявлення та формування трирівневої схеми ризику для пріоритезації подальшого аналізу повідомлень.

Аналіз останніх досліджень і публікацій. У дослідженні Allcott і Gentzkow (2017) розглянуто феномен неправдивих новин у соціальних медіа в контексті політичних процесів. Автори пропонують операційний підхід до ідентифікації неправдивих матеріалів через зовнішню перевірку та демонструють методику емпіричного аналізу на основі корпусу перевірених повідомлень і опитувань щодо розпізнавання та оцінки правдивості. Результат дослідження також показує обмеження оцінювання впливу через неоднорідність аудиторій і специфіку каналів розповсюдження [1].

Огляд Lazer та ін. (2018) формулює наукову постановку проблеми неправдивих новин як сфабрикованої інформації, що імітує новинний формат, але не спирається на редакційні процедури забезпечення точності. Автори підкреслюють необхідність багаторівневих підходів протидії, зокрема на рівні користувача та на рівні інформаційних платформ, і наголошують на потребі доказової оцінки ефективності інтервенцій [2].

Звіт Wardle і Derakhshan (2017) пропонує міждисциплінарну рамку інформаційного безладу, яка описує ключові компоненти процесу поширення неправдивої інформації (суб’єкти, повідомлення,

інтерпретатор) та етапи життєвого циклу (створення, виробництво, дистрибуція). Рамка дозволяє систематизувати індикатори за тим, на якому етапі вони проявляються, і підтримує побудову процедур пріоритизації перевірки та реагування [3].

Емпіричне дослідження Vosoughi, Roy і Aral (2018) на великому масиві даних соціальної платформи показує, що неправдиві повідомлення поширюються далі, швидше та ширше, ніж правдиві, і цей ефект спостерігається в різних тематичних категоріях, з вираженістю у політичному контенті. Окремо аналізуються можливі механізми та демонструється, що відмінності не зводяться лише до активності автоматизованих акаунтів [4].

Систематизацію теорій і методів виявлення неправдивої інформації подано у роботі Zhou і Zafarani (2021), де узагальнено підходи за джерелами сигналів і логікою прийняття рішення. Виділяються контентні індикатори, знаннєво-фактологічні підходи, аналіз поширення та оцінювання надійності джерела, також підкреслюються проблеми перенесення моделей між доменами, змінюваності наративів і вимоги до відтворюваності результатів [5].

Перспектива аналізу даних представлена у роботі Shu та ін. (2017), де задачу виявлення неправдивих новин розглянуто як проблему класифікації з використанням ознак змісту, контексту та поширення. Автори узагальнюють обмеження публічних наборів даних, зокрема різнорідність розмітки, неповноту доступних ознак і наслідки цього для порівнюваності та відтворюваності результатів [6].

Гібридний підхід до виявлення представлено у роботі Ruchansky, Seo і Liu (2017), де модель CSI інтегрує аналіз тексту, характеристики користувачів і джерел, а також часові сигнали взаємодій у процесі поширення. Продемонстровано, що поєднання різних груп сигналів може підвищувати якість виявлення, але якість і застосовність таких підходів залежать від доступності даних про взаємодії та джерела [7].

Контекстно-орієнтований аналіз впливових операцій для України наведено у дослідженні Maschmeyer, Abrahams, Pomerantsev та Yermolenko (2025), де розглядаються межі впливових операцій у соціальних мережах і співвідношення ефектів соціальних платформ та телебачення. Показано контекстну залежність ефектів і відмінності каналів за механізмами охоплення та переконливості, що слід враховувати під час побудови моделей оцінювання ризику та процедур реагування [8].

Безпекова інтерпретація ролі дезінформації у гібридній війні запропонована у роботі Bachmann, Putter і Duczynski (2023), де інформаційні операції розглядаються як інструмент досягнення стратегічних ефектів через підрив довіри та викривлення інформаційного середовища. Акцентовано складність атрибуції та роль створення невизначеності для систем прийняття рішень, що формує вимогу до прикладних підходів, орієнтованих на швидке виявлення і кероване реагування [9].

Робота Тищенка та Мужанової (2022) подає визначення дезінформації як недостовірної, оманливої та маніпулятивної інформації, створеної навмисно з мотивами отримання вигоди, а фейкові новини трактує як один із методів її поширення. Також наведено перелік характерних ознак фейкових новин (маніпулятивність, навмисне введення в оману, використання хибних або анонімних джерел, невідповідність заголовка змісту, імітація легітимних новин, мотиви створення) та узагальнено базові методи виявлення: аналіз джерела, змісту і заголовка, перевірка автора й посилань, перевірка актуальності, фактчекінг, консультація експерта, аналіз емоційної реакції на повідомлення [10].

Метою статті є розроблення методу адаптивного відбору та зважування ознак неправдивої інформації для підвищення ефективності її виявлення в умовах гібридної війни на основі евристичних процедур покрокового формування підмножини ознак, інтегрального скорингу з ваговими коефіцієнтами та порогового тріажу ризику, а також обґрунтування підходів до оцінювання якості за метриками precision, recall, F_1 і PR/ROC-аналізом.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Постановка задачі виявлення неправдивої інформації в текстових повідомленнях розглядається як побудова скорингової функції або класифікатора на основі вектору ознак. Нехай X – множина повідомлень, $y \in \{0,1\}$ – мітка класу, $F = \{f_1, \dots, f_m\}$ – множина кандидатних ознак, а $S \subseteq F$ – підмножина ознак, що використовується для прийняття рішення. Підхід до інтерпретації відбору ознак як пошуку підмножини S, що забезпечує кращу якість за заданим критерієм з урахуванням складності та надлишковості, наведено у дослідженні Kohavi та John (1997) [11].

Базові стратегії відбору ознак у машинному навчанні зручно описувати через поділ на фільтрові та обгорткові підходи. У дослідженні Kohavi та John (1997) фільтрові методи визначаються як оцінювання інформативності ознак незалежно від конкретного алгоритму навчання, тоді як обгорткові методи оцінюють підмножину ознак через якість моделі, навченої на цих ознаках; також зазначається, що нерелевантні або надмірні ознаки можуть погіршувати якість окремих алгоритмів [11].

Критерій якості підмножини ознак доцільно будувати з урахуванням двох вимог: релевантності ознак до класу та низької взаємної надлишковості. У статті Hall (1999) подано підхід Correlation-based Feature Selection, у якому якість підмножини зростає зі збільшенням середньої кореляції ознак з класом і зменшується зі зростанням середньої кореляції між самими ознаками; у роботі наведено формулу оцінювання підмножини через відповідні середні кореляції [12].

Обчислювальна складність повного перебору підмножин зумовлює застосування евристичних процедур пошуку. У статті Hall (1999) описано типові стратегії: нарощення від порожнього набору, скорочення від повного набору та «best-first» пошук із умовою зупинки. Такі стратегії методологічно узгоджуються з покроковими схемами додавання та вилучення ознак під час оптимізації критерію якості [12].

Залежність якості текстової класифікації від вибору метрики ранжування ознак показано у роботі Forman (2003) на масштабному емпіричному порівнянні метрик відбору ознак. У роботі продемонстровано, що різні метрики ранжування можуть давати різну якість, а їхня придатність залежить від властивостей даних і постановки задачі, що обґрунтовує необхідність явного визначення критерію корисності ознак у прикладних методах [13].

Формування контентних ознак тексту зазвичай базується на ваговому поданні термів і нормалізації. У статті Salton та Buckley (1988) описано компоненти вагового подання, включно з частотою термів у документі, інверсною частотою документів та нормалізацією довжини векторів, що є основою для побудови TF-IDF та споріднених ознак у текстових задачах [14].

Статистичні характеристики використовуються для опису структурованості та передбачуваності текстів. У роботі Shannon (1951) введено ентропійний підхід, де ентропія визначається як середня кількість бітів, потрібних на символ, що дозволяє формувати ознаки різноманітності та регулярності текстового потоку як доповнення до контентних індикаторів [15].

Частотно-рангові індикатори лексики можуть бути введені через закономірності природної мови. У статті Piantadosi (2014) узагальнено аналіз закону Ципфа та подано залежність частоти від рангу у вигляді $f(r) \propto 1/r^{\alpha}$, що може використовуватись як теоретичне підґрунтя для ознак, які оцінюють відхилення частотного профілю від типових розподілів [16].

Вибір метрик якості визначається властивостями матриці помилок та практичною ціною помилок різних типів. У дослідженні Sokolova та Lapalme (2009) систематизовано показники якості класифікації та їх інтерпретації, що забезпечує підстави для використання precision, recall та F_1як базових метрик оцінювання якості [17].

Для дисбалансних даних PR-аналіз часто є більш чутливим до якості на позитивному класі. У статті Saito та Rehmsmeier (2015) обґрунтовано, що PR-подання є інформативнішим за ROC-подання в умовах дисбалансу, оскільки безпосередньо відображає співвідношення істинно позитивних рішень серед усіх позитивних передбачень [18].

Зв’язок між PR та ROC формалізовано у роботі Davis та Goadrich (2006), де підкреслено, що оптимізація показників у ROC-просторі не гарантує оптимізації у PR-просторі; також зазначено, що лінійна інтерполяція, коректна для ROC-кривих, не є коректною для PR-кривих. Це слід враховувати при виборі порогів скорингу та при оцінюванні режимів тріажу [19].

МЕТОДИКА ДОСЛІДЖЕННЯ

Дані та формування корпусу. Джерелом даних є відкритий набір українськомовних новин, отриманий з платформи Kaggle [20]. Тексти корпусу тематично пов’язані з подіями війни Російської Федерації проти України та інформаційним порядком денним навколо цих подій. У межах дослідження використано файл news_data.csv, який містить такі поля: Text – повний текст повідомлення; Label – бінарна мітка достовірності зі значеннями True або False; Unnamed: 0 – службовий ідентифікатор запису. Значення Label інтерпретовано так: False відповідає класу неправдивої інформації, True відповідає класу правдивої інформації. На етапі підготовки корпусу вилучено порожні записи та дублікати. Для підвищення змістовності прикладів застосовано фільтрацію за довжиною повідомлення і включено лише тексти обсягом не менше 200 символів.

Попередня обробка текстів. Перед побудовою ознакового опису виконано приведення тексту до нижнього регістру, очищення від надлишкових пробілів і службових символів. Подальше подання текстів здійснювалося в ознаковому просторі на основі частотно-вагового підходу з нормалізацією, що узгоджується з загальноприйнятою практикою термового зважування в задачах аналізу текстів [14].

Ознаковий опис повідомлень. Кожне повідомлення подається як вектор ознак на основі TF-IDF для уніграм і біграм. Для обмеження розмірності та зменшення впливу рідкісних елементів

використовуються стандартні обмеження на мінімальну частоту появи термів та максимальну кількість ознак.

Розбиття даних для навчання та контролю якості. Корпус після очищення та фільтрації розбивається на три частини: навчальну (train), валідаційну (val) та тестову (test). Розбиття виконується стратифіковано зі збереженням пропорції класів. Валідаційна вибірка використовується для вибору розміру підмножини ознак і підбору порогів тріажу. Тестова вибірка використовується лише для підсумкової оцінки якості та не застосовується для підбору параметрів.

Базова модель та скоринг. Як базову модель застосовано лінійний класифікатор логістичної регресії, навчений на TF-IDF ознаках. Для кожного повідомлення модель формує числовий скоринг у діапазоні від 0 до 1, який інтерпретується як оцінка належності повідомлення до класу неправдивої інформації. У процесі навчання враховується дисбаланс класів за рахунок вагової компенсації.

Відбір ознак і адаптивне керування розмірністю. Для зменшення розмірності ознакового простору застосовано евристичну процедуру відбору, що складається з двох кроків: (1) ранжування кандидатних ознак за їх дискримінаційною здатністю на навчальній вибірці, (2) вибір розміру підмножини ознак за результатами перевірки на валідаційній вибірці. Практичну доцільність такого підходу для текстової класифікації, а також чутливість результату до метрики ранжування ознак і розміру підмножини показано у роботі Forman (2003) [13].

Процедура вибору підмножини ознак. Для ранжування ознак застосовано критерій χ^2, після чого перевірено кілька значень K (5000; 10000; 12000; 20000; 30000). Для кожного варіанта перенавчається базова модель на train та оцінюється якість на val за метрикою F_1.Робочий варіант обирається як компроміс між збереженням якості та зменшенням розмірності. У проведеному експерименті обрано $K = 5000$ як мінімальне значення серед перевірених, що забезпечило якість на валідаційній вибірці, співмірну з іншими варіантами [13].

Калібрування порогів і тріаж ризику. На основі скорингу вводиться якісна трирівнева категоризація ризику: низький, середній і високий. Два пороги T_1та T_2 визначаються на валідаційній вибірці за процентилями розподілу скорингу: T_1 відповідає 80-му процентилю, T_2 – 95-му процентилю. Відповідно, до групи високого ризику віднесено верхні 5% повідомлень за скорингом, до групи середнього ризику – наступні 15%, до групи низького ризику – решту 80%. Такий підхід узгоджується з вимогою враховувати різну ціну помилок різних типів при практичному застосуванні [17].

Формальний опис методу. Вхідні дані: розмічений корпус повідомлень, розбитий на train/val/test; параметри побудови TF-IDF ознак; множина значень K для формування підмножин ознак (top-K).

Вихідні дані: підмножина ознак Sрозмірності K^{*}; навчена скоринг-модель $s(x) \in [0; 1]$; пороги T_1і T_2та правило віднесення повідомлення до рівня ризику.

Процедура:

1. Для кожного повідомлення формується TF-IDF подання у просторі уніграм і біграм.
2. На навчальній вибірці виконується ранжування ознак за критерієм χ^2, після чого формуються підмножини S_Kдля заданих значень K.
3. Для кожного Knавчається базова модель на ознаках S_K, а якість оцінюється на валідаційній вибірці за метрикою F_1.
4. Значення K^{*}визначається як таке, що забезпечує максимальне F_1на валідаційній вибірці; у разі рівності значень F_1обирається мінімальне Кяк компроміс між якістю та розмірністю.
5. На підмножині $S = S_{K^{*}}$навчається фінальна модель, яка формує скоринг $s(x)$як оцінку належності повідомлення до класу неправдивої інформації.
6. На основі скорингу застосовується трирівнева схема ризику з використанням порогів T_1і T_2, визначених на валідаційній вибірці.

Метрики якості та принципи оцінювання. Якість базової моделі та моделі після відбору ознак оцінюється на тестовій вибірці за метриками precision, recall та F_1. Додатково застосовується оцінювання за PR-кривою та ROC-кривою. У задачах з дисбалансом класів PR-подання є більш інформативним для оцінювання якості на позитивному класі, що обґрунтовується у статті Saito та Rehmsmeier (2015) [18].

Інтерпретація PR та ROC. Під час аналізу кривих якості враховується, що оптимізація показників у ROC-просторі не гарантує оптимізації у PR-просторі, а також що інтерполяція, коректна для ROC-кривих, не є коректною для PR-кривих. Це враховується при порівнянні режимів порогового рішення та під час підбору порогів тріажу [19].

Відтворюваність. Усі кроки підготовки корпусу, побудови ознак, навчання моделі, відбору ознак, визначення порогів тріажу та підрахунку метрик виконуються програмно в Python і можуть бути відтворені на основі відкритого набору даних та зафіксованих параметрів експерименту. Для відтворення експерименту фіксуються параметри: мінімальна довжина тексту 200 символів, максимальна кількість

TF-IDF ознак 30000, перевірені значення K = {5000, 10000, 12000, 20000, 30000}, пороги тріажу як 80-й та 95-й процентилі скорингу на валідаційній вибірці.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Після вилучення порожніх записів і дублікатів та застосування фільтра довжини повідомлення не менше 200 символів сформовано корпус обсягом 29372 повідомлення. Із них 353 повідомлення належать до класу неправдивої інформації, а 29019 – до класу правдивої інформації. Для оцінювання якості виконано стратифіковане розбиття корпусу на навчальну, валідаційну та тестову вибірки у пропорції 70/15/15, що відповідає 20560, 4406 і 4406 повідомленням відповідно.

Як базову модель використано TF-IDF-подання текстів для уніграм і біграм у поєднанні з лінійним класифікатором логістичної регресії. Після цього проведено відбір ознак за схемою $\chi^2 + top - K$ з перевіркою кількох значень K на валідаційній вибірці. Серед перевірених значень $K = \{5000; 10000; 12000; 20000; 30000\}$ однакове значення $F1$ на валідаційній вибірці було отримано для всіх варіантів, тому обрано мінімальне значення $K = 5000$ як таке, що забезпечує найбільше зменшення розмірності без втрати якості.

Для оцінювання впливу відбору ознак на якість класифікації доцільно порівняти метрики базової моделі та моделі після скорочення ознакового простору. Відповідні результати для валідаційної та тестової вибірок наведено в таблиці 1.

Таблиця 1

Порівняння якості базової моделі та моделі після відбору ознак

Варіант	Набір	Precision	Recall	F1	PR-AUC	ROC-AUC
Baseline (без відбору)	Val	0.708	0.642	0.673	0.727	0.976
	Test	0.826	0.717	0.768	0.819	0.986
Відбір ознак $\chi^2 + top - K$ ($K = 5000$)	Val	0.708	0.642	0.673	0.714	0.974
	Test	0.809	0.717	0.760	0.805	0.984

Наведені результати показують, що після відбору ознак розмірність ознакового простору зменшується з 30000 до 5000 ознак, тобто на 83,3%. При цьому значення F1 на тестовій вибірці зменшується лише з 0.768 до 0.760, а PR-AUC – з 0.819 до 0.805. Отже, зменшення кількості ознак не призводить до суттєвого погіршення якості класифікації і дає змогу зберегти працездатність моделі при значно компактнішому поданні текстів.

Для практичного використання отриманого скорингу застосовано трирівневий тріаж ризику. Пороги $T1$ і $T2$ визначалися на валідаційній вибірці за процентилями розподілу скорингу: $T1$ відповідає 80-му процентилю, $T2$ – 95-му процентилю. Таким чином, до групи високого ризику відносяться верхні 5% повідомлень за скорингом, до групи середнього ризику – наступні 15%, до групи низького ризику – решта 80%. Значення порогів наведено в таблиці 2.

Таблиця 2

Порогові значення для тріажу ризику

Поріг	Значення
T1 (q80)	0.08036
T2 (q95)	0.17705

Після визначення порогів доцільно проаналізувати, як саме повідомлення розподіляються між рівнями ризику на валідаційній і тестовій вибірках, а також скільки фейків концентрується в кожній групі. Відповідний розподіл подано в таблиці 3.

Таблиця 3

Розподіл повідомлень за рівнями ризику

Набір	Рівень ризику	Кількість повідомлень	Кількість фейків	Кількість правдивих
Val	Низький	3524	2	3522
	Середній	661	7	654
	Високий	221	44	177
Test	Низький	3451	1	3450
	Середній	699	2	697
	Високий	256	50	206

На тестовій вибірці високоризикова група охоплює 256 повідомлень із 4406, тобто близько 5,8% потоку, але містить 50 із 53 фейків тестової вибірки. Водночас у низькоризиковій групі, що охоплює 3451 повідомлення, зафіксовано лише 1 фейк. Це свідчить про те, що запропонована схема тріажу концентрує більшість неправдивих повідомлень у невеликій частині потоку, яка може бути визначена як пріоритетна для ручної перевірки.

Отримані результати підтверджують працездатність запропонованого підходу в двох аспектах. По-перше, застосування відбору ознак дає змогу істотно скоротити розмірність ознакового простору без суттєвої втрати якості класифікації. По-друге, використання трирівневого тріажу на основі процентилів скорингу формує практично придатну схему пріоритезації повідомлень, у межах якої більшість фейків концентрується у високоризиковій групі.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

На тестовій вибірці високоризикова група охоплює 256 повідомлень із 4406, тобто близько 5,8% потоку, але містить 50 із 53 фейків тестової вибірки. Водночас у низькоризиковій групі, що охоплює 3451 повідомлення, зафіксовано лише 1 фейк. Це свідчить про те, що запропонована схема тріажу концентрує більшість неправдивих повідомлень у невеликій частині потоку, яка може бути визначена як пріоритетна для ручної перевірки.

Отримані результати підтверджують працездатність запропонованого підходу в двох аспектах. По-перше, застосування відбору ознак дає змогу істотно скоротити розмірність ознакового простору без суттєвої втрати якості класифікації. По-друге, використання трирівневого тріажу на основі процентилів скорингу формує практично придатну схему пріоритезації повідомлень, у межах якої більшість фейків концентрується у високоризиковій групі.

У статті розроблено метод адаптивного відбору та зважування ознак неправдивої інформації для підвищення ефективності її виявлення в умовах гібридної війни. Метод поєднує контентне TF-IDF-подання текстів, евристичне скорочення ознакового простору за схемою ранжування і вибору підмножини ознак, побудову скорингової оцінки належності повідомлення до класу неправдивої інформації та трирівневий тріаж ризику на основі порогів, визначених за процентилями скорингу.

Експериментальну перевірку виконано на відкритому українськомовному корпусі новин, пов’язаних із тематикою війни Російської Федерації проти України. Після очищення корпусу та застосування фільтра довжини сформовано вибірку обсягом 29372 повідомлення. Порівняння базової моделі та моделі після відбору ознак показало, що скорочення ознакового простору з 30000 до 5000 ознак не призводить до суттєвого погіршення якості: на тестовій вибірці значення F1 зменшується лише з 0.768 до 0.760, а PR-AUC – з 0.819 до 0.805. Це свідчить про доцільність застосування відбору ознак для зменшення розмірності представлення текстів без істотної втрати ефективності виявлення.

Окремим результатом є підтвердження практичної придатності тріажу ризику. За використання порогів, визначених як 80-й і 95-й процентилі скорингу на валідаційній вибірці, високоризикова група на тестовій вибірці охопила 256 повідомлень із 4406 і містила 50 із 53 фейків, тоді як у низькоризиковій групі зафіксовано лише 1 фейк. Отриманий результат показує, що запропонований підхід придатний не лише для класифікації повідомлень, а й для пріоритезації їх перевірки в умовах обмежених ресурсів.

Практичне значення запропонованого методу полягає в можливості використання його як інтерпретованого та ресурсно-ефективного компонента в системах моніторингу інформаційного простору. Перспективи подальших досліджень пов’язані з розширенням набору ознак за рахунок семантичних, джерельних і поведінкових індикаторів, перевіркою методу на інших українськомовних корпусах, а також із дослідженням більш складних схем адаптивного зважування ознак і калібрування порогів ризику.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211-236. https://doi.org/10.1257/jep.31.2.211
2. Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

3. Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html
4. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
5. Zhou, X., & Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), Article 109, 1-40. https://doi.org/10.1145/3395046
6. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22-36. https://doi.org/10.1145/3137597.3137600
7. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797-806). ACM. https://doi.org/10.1145/3132847.3132877
8. Maschmeyer, L., Abrahams, A., Pomerantsev, P., & Yermolenko, V. (2025). Donetsk don't tell – "hybrid war" in Ukraine and the limits of social media influence operations. *Journal of Information Technology & Politics*, *22*(1), 49-64. https://doi.org/10.1080/19331681.2023.2211969
9. Bachmann, S.-D. D., Putter, D., & Duczynski, G. (2023). Hybrid warfare and disinformation: A Ukraine war perspective. *Global Policy*, *14*(5), 858-869. https://doi.org/10.1111/1758-5899.13257
10. Tyshchenko, V. S., & Muzhanova, T. M. (2022). Dezinformatsiia i feikovi novyny: Oznaky ta metody vyiavlennia v merezhi Internet [Disinformation and fake news: Features and methods of detection on the Internet]. *Kiberbezpeka: osvita, nauka, tekhnika*, *2*(18), 175-186. https://doi.org/10.28925/2663-4023.2022.18.175186
11. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X
12. Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, University of Waikato). https://www.cs.waikato.ac.nz/ml/publications/1999/99MH-Thesis.pdf
13. Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289-1305. https://www.jmlr.org/papers/v3/forman03a.html
14. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0
15. Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50-64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x
16. Piantadosi, S. T. (2014).Zipf's word frequency law in natural language:A critical review and future directions.*Psychonomic Bulletin & Review*, *21*(5), 1112-1130. https://doi.org/10.3758/s13423-014-0585-6
17. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002
18. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
19. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). ACM. https://doi.org/10.1145/1143844.1143874
20. Zepopo. (n.d.). *Ukrainian fake and true news* [Data set]. Kaggle. https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023

Dmytro Dzhendzhero
Postgraduate Student of the Department of Cybersecurity and Information Protection,
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID:0009-0007-9999-850X
dzhendzherod@gmail.com

Volodymyr Nakonechnyi
Doctor of Technical Sciences, Professor,
Professor of the Department of Cybersecurityand Information Protection
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: 0000-0002-0247-5400
volodym.nakonechnyi@knu.ua

METHOD OF ADAPTIVE SELECTION AND WEIGHTING OF FALSE INFORMATION INDICATORS TO IMPROVE DETECTION EFFICIENCY UNDER HYBRID WARFARE CONDITIONS

Abstract. The article addresses the problem of improving false information detection in text messages under hybrid warfare conditions. The relevance of the study is determined by the fact that disinformation campaigns in the modern information environment are used to undermine trust in public institutions, distort the perception of events, destabilize public attitudes, and create additional pressure on decision-making systems. The introduction substantiates the need for an interpretable and resource-efficient method suitable for large, dynamic, and imbalanced message streams. The problem statement shows that the use of the full feature space increases computational complexity, reduces interpretability, and does not support the prioritization of message verification according to risk level. The section on recent studies and publications summarizes current approaches to false information detection, including content-based, fact-based, behavioral, and hybrid models, as well as approaches to disinformation risk assessment in wartime conditions. The theoretical foundations systematize the main concepts related to feature selection, text feature space construction, term weighting, classification quality assessment, and the use of PR/ROC representations under class imbalance. On this basis, the conceptual framework of the proposed method of adaptive selection and weighting of false information indicators is formed. The methodology section describes an experimental evaluation carried out on an open Ukrainian-language news corpus related to the events of Russia's war against Ukraine. After cleaning the data, removing empty records and duplicates, and applying a minimum length filter of 200 characters, a dataset of 29,372 messages was obtained, including 353 false messages and 29,019 true messages. TF-IDF features based on unigrams and bigrams were used for text representation, while logistic regression was selected as the baseline classifier. Feature selection was implemented using a chi-square plus top-K scheme with several K values tested on the validation set; the final working configuration was K=5000. For practical decision support, a three-level risk triage scheme was introduced based on the 80th and 95th percentiles of the validation score distribution. The results section shows that reducing the feature space from 30,000 to 5,000 features does not lead to a substantial decrease in classification quality: on the test set, F1 decreases only from 0.768 to 0.760, while PR-AUC decreases from 0.819 to 0.805. At the same time, the proposed triage procedure demonstrates practical value: the high-risk group covered 256 messages out of 4,406 in the test set and contained 50 out of 53 false messages, whereas the low-risk group contained only 1 false message. The conclusions justify that the proposed method can be used as an interpretable and resource-efficient component of information space monitoring systems. Further research should focus on extending the feature set with semantic, source-based, and behavioral indicators, as well as testing the method on additional Ukrainian-language corpora.

Keywords: false information; fake news detection; hybrid warfare; feature selection; TF-IDF; logistic regression; risk triage; information security.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211-236. https://doi.org/10.1257/jep.31.2.211

2. Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998
3. Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html
4. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
5. Zhou, X., & Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys, 53*(5), Article 109, 1-40. https://doi.org/10.1145/3395046
6. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36. https://doi.org/10.1145/3137597.3137600
7. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797-806). ACM. https://doi.org/10.1145/3132847.3132877
8. Maschmeyer, L., Abrahams, A., Pomerantsev, P., & Yermolenko, V. (2025). Donetsk don't tell – "hybrid war" in Ukraine and the limits of social media influence operations. *Journal of Information Technology & Politics, 22*(1), 49-64. https://doi.org/10.1080/19331681.2023.2211969
9. Bachmann, S.-D. D., Putter, D., & Duczynski, G. (2023). Hybrid warfare and disinformation: A Ukraine war perspective. *Global Policy, 14*(5), 858-869. https://doi.org/10.1111/1758-5899.13257
10. Tyshchenko, V. S., & Muzhanova, T. M. (2022). Dezinformatsiia i feikovi novyny: Oznaky ta metody vyiavlennia v merezhi Internet [Disinformation and fake news: Features and methods of detection on the Internet]. *Kiberbezpeka: osvita, nauka, tekhnika, 2*(18), 175-186. https://doi.org/10.28925/2663-4023.2022.18.175186
11. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X
12. Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, University of Waikato). https://www.cs.waikato.ac.nz/ml/publications/1999/99MH-Thesis.pdf
13. Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289-1305. https://www.jmlr.org/papers/v3/forman03a.html
14. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0
15. Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30*(1), 50-64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x
16. Piantadosi, S. T. (2014).Zipf's word frequency law in natural language:A critical review and future directions.*Psychonomic Bulletin & Review, 21*(5), 1112-1130. https://doi.org/10.3758/s13423-014-0585-6
17. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002
18. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
19. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). ACM. https://doi.org/10.1145/1143844.1143874
20. Zepopo. (n.d.). *Ukrainian fake and true news* [Data set]. Kaggle. https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news

Отримано редакцією журналу / Received: 30.01.26
Прорецензовано / Revised: 12.02.26
Схвалено до друку / Accepted: 25.06.26