Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

УДК 004.8 DOI: https://doi.org/10.30837/0135-1710.2026.189.153

Даценко С. С., Кучук Г. А.

МІЖВИБІРКОВА УЗАГАЛЬНЕНІСТЬ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН ІЗ ВИКОРИСТАННЯМ КОНТРАСТИВНОГО АДАПТАЦІЙНОГО НАВЧАННЯ

Потреби сучасного промислового бізнесу дедалі частіше стикаються з викликами, пов'язаними з поширенням фейкових новин, які можуть суттєво впливати на репутацію, довіру клієнтів і фінансові показники промислових компаній. Наявні системи виявлення демонструють майже ідеальну точність на окремих еталонних наборах даних, проте зазнають катастрофічної невдачі за умови застосування до інформації з різних джерел, що викликає серйозні сумніви щодо готовності до впровадження в реальних умовах. **Предмет дослідження** – методи виявлення фейкових новин у структурно різноманітних наборах даних. **Метою** є розроблення й валідація підходу до виявлення фейкових новин у наборах даних зі структурно різноманітним вмістом із використанням контрастивного адаптаційного навчання, що дасть змогу створити єдину модель на основі трансформера. **Завдання дослідження:** встановлення базових показників для окремих наборів даних за допомогою мовної моделі DeBERTa-v3-base на трьох різних наборах даних фейкових новин; побудова систематичної матриці перенесення між наборами даних для кількісного оцінювання невдач узагальнення; надання пропозиції та оцінювання контрастивного об'єднаного адаптаційного навчання як підходу адаптації до домену; дослідження з абляцією контрастивних гіперпараметрів. **Метод дослідження.** Мовна модель DeBERTa-v3-base налаштовується на трьох наборах даних фейкових новин ISOT, LIAR та WELFake. Комбінована навчальна задача інтегрує класифікацію за крос-ентропією з контрольованою контрастивною втратою по об'єднаних наборах даних з використанням збалансованого вибіркового відбору з різних наборів даних. **Результати.** Матриця перенесення між наборами даних свідчить про катастрофічну невдачу узагальнення: модель, яка досягає 100% F1 на ISOT, демонструє лише 23,9% на WELFake та 40,3% на LIAR. Запропонована контрастивна об'єднана модель створює єдину модель, яка одночасно досягає 97,5% F1 на ISOT, 57,4% на LIAR і 98,4% на WELFake з абсолютним покращенням F1 до 97,9% порівняно з базовими переносами**. Висновки.** Абляційний аналіз підтверджує, що контрастивна задача має особливу користь у складніших доменах. Контрастне об'єднане адаптаційне навчання створює моделі, придатні для розгортання в середовищах новин із декількома джерелами.

Ключові слова: виявлення фейкових новин; міжвибіркова узагальненість; трансформерна мовна модель; трансферне навчання; оброблення природної мови.

Вступ

У сучасному інформаційному середовищі промисловий бізнес дедалі частіше стикається з викликами, пов'язаними з поширенням фейкових новин, які можуть суттєво впливати на репутацію, довіру клієнтів і фінансові показники промислових компаній. У відповідь на ці загрози формується цілий спектр інструментів для виявлення й нейтралізації дезінформації, що активно інтегруються в корпоративні процеси [1]. Компанії впроваджують системи моніторингу медіа, які дають змогу в реальному часі відстежувати згадки про бренд і оперативно реагувати на підозрілий контент. Важливе значення відіграють технології штучного інтелекту, зокрема рішення на основі Natural Language Processing, які аналізують тональність, структуру й достовірність текстів. Окрім цього, використовуються інструменти аналізу соціальних мереж, що допомагають виявляти аномальні патерни поширення інформації та координовані кампанії дезінформації [2]. Значну увагу приділяють також автоматизованим системам оцінювання ризиків, що сигналізують про потенційні загрози для бренду ще до того, як вони набудуть широкого розголосу [3]. Бізнес активно співпрацює із зовнішніми платформами й експертними організаціями, щоб підвищити точність перевірки фактів. Водночас інтеграція таких інструментів у внутрішні комунікаційні стратегії дає змогу підвищити рівень інформаційної стійкості компаній. Не менш важливим є використання систем управління репутацією, які поєднують аналітику, автоматизацію та кризовий менеджмент [4]. Усе це формує новий підхід до корпоративної безпеки, де інформаційні ризики розглядаються нарівні з фінансовими чи операційними. Отже, інструменти для виявлення фейкових новин стають невід'ємною частиною стратегічного управління бізнесом і логічно продовжують ширший контекст трансформацій інформаційного середовища.

ISSN 3083-7715 (online) Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)
ISSN 3083-7650 (0135-1710) (print) Management Information System and Devices. 2026. No. 2 (189)

Методи машинного навчання продемонстрували значні успіхи у виявленні фейкових новин під час оцінювання на окремих наборах даних. Традиційні алгоритми, як-от LightGBM і XGBoost, досягають показників F1, що перевищують 99% на широко використовуваному наборі даних ISOT [5], тоді як моделі глибокого навчання, зокрема BERT і RoBERTa, повідомляють про аналогічно ефективні результати [6]. У роботі підтверджено ці висновки й продемонстровано, що LightGBM з функціями TF-IDF досягає 99,8% F1 на ISOT, перевершуючи складніші архітектури глибокого навчання за умов обмежених ресурсів [5]. Подальше дослідження гібридних нейронних архітектур (BiLSTM-CNN, BiLSTM-GRU, BiLSTM-GNN) на наборі даних FEVER досягло точності 79,5% зі значними проблемами перенавчання [7].

Однак високі результати на окремих наборах даних створюють оманливу уяву про можливості виявлення. Моделі, навчені на одному наборі даних, вивчають статистичні закономірності, специфічні для конструкції цього набору даних (стилі письма, властиві для джерела, тематичні упередження й артефакти колекції), а не узагальнювальні показники обману [8]. Унаслідок розгортання в реальних умовах, де новинний контент різко різниться за джерелом, стилем, довжиною та предметною галуззю, ці моделі зазнають катастрофічних невдач. Це фундаментальне обмеження залишилося без достатньої уваги в сучасній літературі, до того ж переважну більшість досліджень з виявлення фейкових новин проведено тільки на одному наборі даних [9].

1 Огляд літератури й формулювання проблеми

1.1 Традиційні підходи й підходи на основі глибокого навчання

Галузь автоматизованого виявлення фейкових новин швидко розвивалася, проходячи шлях від методів, що ґрунтуються на правилах, традиційного машинного навчання до складних архітектур глибокого навчання. Ранні підходи, основані на контенті, спиралися на лінгвістичну інженерію ознак – аналіз стилю письма, метрик читабельності, шаблонів пунктуації та індикаторів настрою [10]. Ці вручну створені ознаки в поєднанні з методами градієнтного підсилення, як-от LightGBM і XGBoost, довели надзвичайну ефективність на еталонних наборах даних, часто перевершуючи більш складні нейронні архітектури [5, 11].

Підходи, основані на глибокому навчанні, дали змогу автоматично виявляти відповідні ознаки в необробленому тексті. Рекурентні архітектури (LSTM, BiLSTM, GRU) беруть до уваги послідовні залежності [12], тоді як згорткові нейронні мережі виокремлюють локальні шаблони *n*-грам [13]. Гібридні моделі, що поєднують ці парадигми, продемонстрували конкурентоспроможну продуктивність [7, 14]. Поява попередньо навчених моделей *Transformer* ознаменувала зміну парадигми: BERT [15], RoBERTa [16] і DeBERTa [17] досягають найсучасніших результатів завдяки двоспрямованому контекстному розумінню, водночас механізми розплетеної уваги DeBERTa досягають кращої продуктивності з меншою кількістю параметрів, ніж BERT-large.

1.2 Оцінювання на різних наборах даних і зсув домену

Незважаючи на високі результати на окремих наборах даних, узагальнення між наборами даних залишається критично нерозв'язаним питанням. Проблема зсуву домену, коли розподіл даних навчання $P_{source}(X,Y)$ принципово відрізняється від цільового розподілу $P_{\text{target}}(X,Y)$, гостро відчувається у виявленні фейкових новин, оскільки різні набори даних відтворюють різні аспекти обману [8].

Автори публікації [18] провели одне з небагатьох систематичних оцінювань наборів даних і виявили, що в моделях, навчених на ISOT, падіння точності становило 37–42% за умови тестування на інших наборах даних, до того ж моделі *FakeNewsNet* досягли лише випадкової продуктивності на невидимих даних. Нещодавно Груенштейдл і Кіранне [19] провели перехресне оцінювання даних і моделей, використовуючи попередньо навчені моделі сімейства BERT на європейських наборах даних новин, і підтвердили значне зниження продуктивності на невидимих доменах. Однак це дослідження було суто діагностичним – автори оцінювали наявні моделі без пропозиції методів покращення перехресного узагальнення наборів даних. Описаний в роботі [19] експеримент розширює цей напрям завдяки побудові контрольованої матриці передачі з єдиною архітектурою та впровадженню контрастивного об'єднаного адаптаційного навчання як конкретного рішення.

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

Кілька факторів впливають на погане перенесення між наборами даних:

– упередженість джерела, коли справжні та фейкові новини взяті із систематично різних джерел;

– тематична упередженість, коли навчальні дані охоплюють обмежені предметні галузі;

– структурна упередженість, коли набори даних принципово різняться за текстовими характеристиками.

1.3 Контрастивне навчання для адаптації домену

Контрастивне навчання стало потужною технікою для навчання представлень, стійкого до змін у розподілі даних. Фреймворк SimCLR [20] продемонстрував високу переносимість навчання ознак, тоді як *Supervised Contrastive Learning* (SupCon) [21] розширив ці можливості, даючи змогу використовувати інформацію про мітки. У NLP контрастивні цілі застосовувалися до вбудовування речень [22], класифікації за кількома спробами [23] й адаптації домену [24]. Основним вагомим висновком є те, що контрастивна втрата порівняння заохочує модель вивчати незалежні від домену ознаки способом побудови позитивних пар зі зразків, що мають однакову мітку в різних наборах даних.

Тоді як наявні методи контрастивного навчання для виявлення фейкових новин зосереджені на крос-модальному узгодженні між текстом і зображеннями [25, 26], такий підхід застосовує порівняльне навчання до адаптації доменів між наборами даних, що є принципово іншою метою.

2 Мета й завдання дослідження

Предметом дослідження є методи виявлення фейкових новин у структурно різноманітних наборах даних. Метою – розроблення й валідація підходу до виявлення фейкових новин у наборах даних зі структурно різноманітним вмістом із використанням контрастивного адаптаційного навчання, що дасть змогу створити єдину модель на основі трансформера. У такий спосіб буде подоланий розрив між високими показниками на еталонних наборах даних і готовністю до впровадження в реальних умовах.

Для досягнення окресленої мети необхідно виконати такі завдання:

1) встановити базові показники DeBERTa-v3-base для трьох структурно різноманітних наборів даних фейкових новин (ISOT, LIAR, WELFake) з метою кількісного оцінювання меж продуктивності для окремих наборів даних;

2) побудувати систематичну матрицю перенесення між наборами даних для кількісного оцінювання невдач у генералізації;

3) запропонувати й оцінити контрастивне об’єднане адаптаційне навчання як підхід адаптації до домену;

4) дослідити абляцію контрастивних гіперпараметрів (вага втрат і температура) для визначення оптимальних конфігурацій.

3 Методологія

3.1 Набори даних

Для перевірки запропонованого підходу використовувались три структурно різноманітні набори даних, що подають різні аспекти проблеми виявлення фейкових новин.

Набір даних ***ISOT Fake News Dataset*** [27] містить 44 889 статей зі справжньою інформацією, отриманою з *Reuters*, і фейковими новинами з ненадійних джерел, позначеними організаціями з перевірки фактів. Статті мають середню довжину – 2252 символи зі збалансованими бінарними мітками (47,7% реальні, 52,3% фейкові).

Набір даних ***LIAR Dataset*** [28] містить 12 789 стислих політичних заяв від *PolitiFact*. Оригінальні мітки з шести класів бінаризовані (pants-fire + false + badly-true → fake; half-true + mostly-true + true → real). Твердження мають середню довжину – лише 134 символи, що робить це принципово іншим завданням, яке вимагає розуміння правдивості політичних тверджень, а не стилістичного аналізу. Набір даних ***WELFake Dataset*** [29] містить новинні статті, об’єднані з чотирьох джерел: *Kaggle, McIntire, Reuters* і *BuzzFeed Political*. Оскільки у WELFake є статті *Reuters*, що перетинаються зі справжніми новинами ISOT, тому було виконано систематичну дедуплікацію з використанням 200-символьних текстових відбитків і вилученням 38 285 статей, що перетинаються. Дедуплікований набір даних містить 25 346 статей (46,4% справжніх, 53,6% фейкових) із середньою довжиною 34 56 символів. Усі набори даних були попередньо оброблені за допомогою уніфікованого конвеєра: вилучення URL-адрес і електронної пошти, очищення шаблонів *Reuters*, нормалізація пробілів і фільтрація мінімальної довжини (20 символів). Кожен набір

ISSN 3083-7715 (online) Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)
ISSN 3083-7650 (0135-1710) (print) Management Information System and Devices. 2026. No. 2 (189)

даних був розділений на навчальний, валідаційний і тестовий набори за допомогою стратифікованого розподілу 70/15/15 з фіксованим випадковим початковим числом.

У табл. 1 подано основні характеристики трьох наборів даних після попереднього оброблення.

Таблиця 1. *Характеристики набору даних після попереднього оброблення*

Набір даних	Зразки	Класи	Середня довжина	Вихідний ресурс
ISOT	44,889	2	2,252 chars	Reuters + flagged sources
LIAR (binary)	12,789	2	134 chars	PolitiFact statements
WELFake	25,346	2	3,456 chars	Multi-source (deduplicated)

3.2 Мовна модель DeBERTa-v3-base

Було використано DeBERTa-v3-base [17] як базову мовну модель. Трансформерна модель DeBERTa запроваджує розплутану увагу, подаючи кожен токен за допомогою окремих векторів вмісту й позиції, а також удосконалений декодер з маскою, який зважає на абсолютну позицію в шарі декодування. Модель DeBERTa-v3 додатково підвищує ефективність завдяки попередньому навчанню з виявленням замінених токенів у стилі ELECTRA. Базова модель містить приблизно 86 млн параметрів із прихованим виміром 768 та 12 шарами *Transformer*. Представлення токенів [CLS] є агрегованим представленням послідовності для класифікації.

Для стандартного адаптаційного навчання (експерименти 1–2) архітектура моделі містить базову мережу DeBERTa, шар відсіву $(p = 0{,}1)$ та лінійну класифікаційну головку, навчену із втратою перехресної ентропії.

Для навчання на перехресних наборах даних (експеримент 3) архітектура була розширена за допомогою проєкційної головки для контрольованого контрастивного навчання. Представлення [CLS] надходить до двох паралельних головок: головки класифікації (768 → num_classes, втрата перехресної ентропії) та головки проєкції (768 → 256 → 128, LayerNorm + ReLU, L2-нормалізовані вбудовування для втрат SupCon).

Контрастивна втрата з наставником [21] визначається як

$$L_{SupCon} = \sum_i \frac{-1}{\left|P(i)\right|} \sum_{p \in P(i)} \log \left(\frac{\exp\left(\frac{z_i \cdot z_p}{\tau} \right)}{\sum_{a \neq i} \left(\frac{z_i \cdot z_a}{\tau} \right)} \right),$$

де $P(i)$ – набір зразків, що мають ту саму мітку, що й анкер i; z – L2-нормалізовані проєкції; τ – параметр температури. Загальна мета полягає в $L_{total} = L_{CE} + \beta \cdot L_{SupCon}$, де β контролює вагу контрастиву. Найважливіше те, що позитивні пари будуються через межі наборів даних, що змушує модель навчатися представляти класи, незалежно від домену.

3.3 Реалізація підходу виявлення фейкових новин із використанням контрастивного адаптаційного навчання

У запропонованому підході реалізовано алгоритм збалансованого пакетного відбору, який будує кожну партію з приблизно рівними частинами від кожної комбінації (набір даних, мітка), забезпечуючи позитивні пари між наборами даних у кожній партії.

У всіх проведених експериментах використовувався оптимізатор AdamW із косинусною кривою зміни швидкості навчання й навчанням із змішаною точністю. Повний набір гіперпараметрів подано в табл. 2.

Апаратне забезпечення, що використовувалося в експериментах, мало такі характеристики: NVIDIA GeForce RTX 5070 Ti (16 GB VRAM), Windows 11.

Загалом було визначено чотири типи проведених експериментів, призначених для поетапного оцінювання проблеми узагальнення й запропонованого рішення.

В експерименті типу 1 модель DeBERTa-v3-base проходить незалежне адаптаційне навчання на кожному наборі даних для встановлення базових показників ефективності в межах домену, зокрема з 6-класовою оцінкою LIAR для аналізу детальної класифікації правдивості.

В експерименті типу 2 кожна модель з експерименту 1 оцінюється на тестових наборах усіх інших бінарних наборів даних без додаткового навчання, створюючи матрицю перенесення 3×3, яка кількісно оцінює узагальнення між наборами даних.

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

Таблиця 2. *Гіперпараметри навчання*

Параметр	Значення
Оптимізатор	AdamW
Швидкість навчання	2×10^{-5}
Зниження ваги	0.01
Графік LR	Косинусний з 10% розігрівом
Розмір партії	16
Максимальна кількість епох	5 (3 для абляції)
Терпіння до раннього зупинення	2
Змішана точність	fp16
Максимальна довжина послідовності	512 (256 для абляції)
Контрастивна вага (β)	0.5
Температура (τ)	0.07
Розмірність проєкції	128

В експерименті типу 3 всі три бінарні набори даних об'єднуються для спільного навчання з контрастивною архітектурою, де збалансований вибірник забезпечує наявність позитивних пар між наборами даних у кожному пакеті.

Експеримент типу 4 орієнтований на проведення систематичної абляції над контрастивною вагою $\beta \in \{0,0;\ 0,3;\ 0,5;\ 1,0\}$ і температурою $\tau \in \{0,05;\ 0,07;\ 0,2\}$, де $\beta = 0,0$ – об'єднане навчання лише за CE без контрастивної мети, ізолюючи внесок об'єднання даних від контрастивного сигналу.

Для доповнення кількісного аналізу вбудовування токенів [CLS] як з базової, так і з контрастної моделей проєктуються у два виміри за допомогою t-SNE для якісного аналізу навчених просторів представлення. Усі експерименти аналізують точність, прецизію, відкликання, показник F1 та AUC-ROC, до того ж F1 слугує основним метричним показником порівняння.

Як приклад, що підтверджує ефективність запропонованого підходу, наведено пряме порівняння: модель, навчена виключно на наборі ISOT, досягає показника F1 на рівні 0,5% за умови перенесення на набір WELFake, тоді як контрастивна об'єднана модель, навчена на тих самих даних, але з використанням запропонованої багатозадачної мети, досягає показника F1 на рівні 98,4% на наборі WELFake, що становить абсолютне поліпшення на 97,9 процентних пунктів порівняно з єдиною уніфікованою моделлю.

4 Результати моделювання

Експерименти типу 1 проводилися з метою отримання базових показників для кожного окремого набору даних. У табл. 3 наведено результати тестування за п'ятьма визначеними параметрами в межах домену для моделі DeBERTa-v3-base, налаштованої окремо на кожному наборі даних.

Таблиця 3. *Базові результати для одного набору даних*

Набір даних	Точність	Прецизія	Відтворюваність	F1	AUC-ROC
ISOT	1.000	1.000	1.000	1.000	1.000
WELFake	0.995	0.996	0.996	0.996	1.000
LIAR (binary)	0.590	0.611	0.205	0.307	0.633
LIAR (6-class)	0.205	0.042	0.205	0.070	0.496

Результати проведених експериментів типу 1 демонструють суттєву різницю в продуктивності на різних наборах даних. Так, датасети ISOT і WELFake досягають майже ідеальної класифікації: DeBERTa отримує 100% балів на ISOT і 99,5% на WELFake, що узгоджується з попередніми результатами, відповідно до яких LightGBM досягає 99,8% F1 на ISOT [5]. Ці тестові

ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

набори фактично досягли реальної межі можливостей.

На противагу цьому, бінарний алгоритм LIAR досягає лише 59,0% точності з 30,7% F1, що демонструє невід'ємну складність визначення правдивості тверджень із стислих політичних заяв (медіана 134 символи). Варіант із 6 класами має точність 20,5% (трохи вище 16,7% випадкової базової лінії). У цьому разі модель деградує до прогнозування одного класу більшості. Ці результати підтверджують, що фейкові новини залишаються невиявленими навіть за допомогою найсучасніших трансформерів.

Унаслідок проведення експериментів типу 2 було сформовано матрицю перенесення між наборами даних, що є основним емпіричним внеском цього дослідження. Результати експериментів узагальнено в табл. 4 та подано у вигляді теплової карти на рис. 1. У матриці перенесення між наборами даних наведені на основній діагоналі узагальнені результати оцінки F1 демонструють продуктивність у межах домену.

Таблиця 4. *Матриця перенесення між наборами даних (оцінка F1)*

Тренування / тест	ISOT	LIAR	WELFake
ISOT	**1.000**	0.403	0.239
LIAR	0.057	**0.307**	0.005
WELFake	0.105	0.522	**0.996**

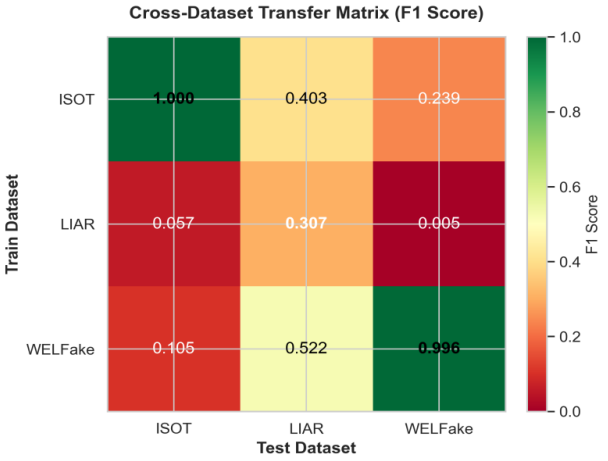


Рис. 1. Теплова карта матриці перенесення між наборами даних

Отримана узагальнена матриця перенесення свідчить про катастрофічну невдачу узагальнення. Показник F1 моделі ISOT знижується зі 100% до 40,3% на наборі LIAR і до 23,9% на наборі WELFake. Модель LIAR дає 5,7% F1 на ISOT і 0,5% на WELFake – гірше, ніж випадковий вибір. Погіршення якості перенесення коливається від 47,5% (WELFake→LIAR) до 98,5% (LIAR→WELFake).

У процесі проведення експериментів типу 3 здійснено адаптаційне навчання контрастивної об'єднаної моделі. У табл. 5 подано узагальнені результати щодо продуктивності контрастивної об'єднаної моделі, оціненої для кожного набору даних окремо.

Таблиця 5. *Продуктивність контрастивної об'єднаної моделі*

Набір даних	Точність	Прецизія	Відтворюваність	F1	AUC-ROC
ISOT	0.974	0.998	0.953	0.975	1.000
LIAR (binary)	0.639	0.601	0.549	0.574	0.672
WELFake	0.982	0.978	0.990	0.984	0.997

Контрастивна об'єднана модель демонструє значне покращення продуктивності порівняно з усіма базовими моделями перенесення між наборами даних (див. рис. 2). Найбільш разючі покращення спостерігаються саме там, де базові моделі зазнавали найбільших невдач: LIAR→WELFake покращується з 0,005 до 0,984 (+97,9% абсолютного F1), LIAR→ISOT – з 0,057 до 0,975 (+91,8%), а WELFake→ISOT – з 0,105 до 0,975 (+87,0%). Ця єдина модель зберігає високу ефективність у межах домену: 97,5% F1 на ISOT (−2,5% порівняно зі спеціалізованим базовим показником), 98,4% на WELFake (−1,2%) та 57,4% на LIAR (+26,7 процентних пунктів порівняно з базовим показником для одного набору даних).

Експерименти типу 4 були орієнтовані на дослідження абляції. Абляція в аналізі фейкових новин є важливим методом оцінювання впливу окремих ознак на ефективність моделі. Вона передбачає послідовне вилучення певних компонентів, як-от текстові характеристики, джерело інформації чи заголовок, щоб визначити їх значущість. Завдяки цьому дослідники можуть зрозуміти, які саме фактори найбільше впливають на точність виявлення неправдивих повідомлень. Абляційний аналіз також дає змогу встановити зайві або малоефективні ознаки,

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

що не мають суттєвого впливу на результат. Це сприяє оптимізації моделей, роблячи їх простішими, швидшими та більш інтерпретованими. Отже, використання абляції підвищує прозорість і надійність систем автоматичного розпізнавання фейкових новин.

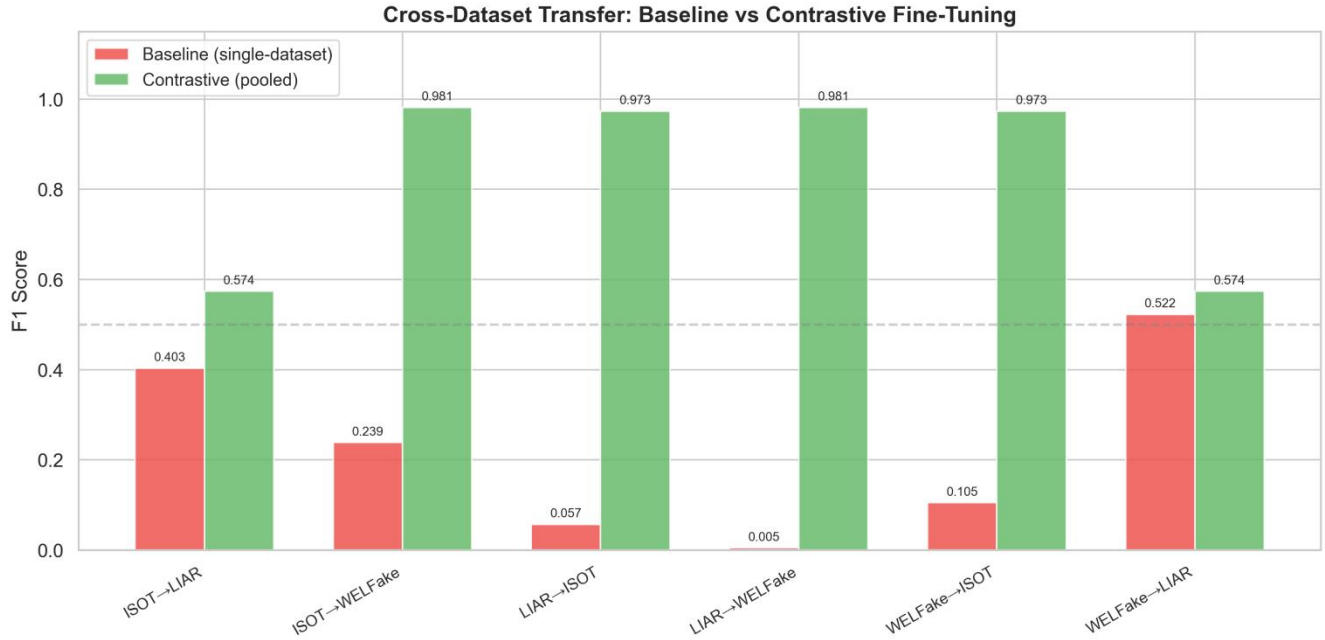


Рис. 2. Порівняння показника F1 між наборами даних для всіх пар перенесення: червоні стовпці позначають базові результати перенесення для одного набору даних; зелені стовпці – результати контрастивної об'єднаної моделі

У табл. 6 і 7 подано результати систематичної абляції залежно від контрастного коефіцієнта β і температури τ відповідно.

Таблиця 6. *Абляція: контрастна вага β (оцінки F1)*

β	ISOT	LIAR	WELFake	Середнє
0.0 (CE only)	0.973	0.585	0.978	0.845
0.3	0.971	0.573	0.978	0.841
0.5	0.975	0.607	0.973	0.852
1.0	0.973	0.574	0.980	0.842

Таблиця 7. *Абляція: температура τ (оцінки F1, де β = 0,5)*

τ	ISOT	LIAR	WELFake	Середнє
0.05	0.972	0.566	0.975	0.838
0.07	0.974	0.566	0.977	0.839
0.20	0.974	0.606	0.974	0.851

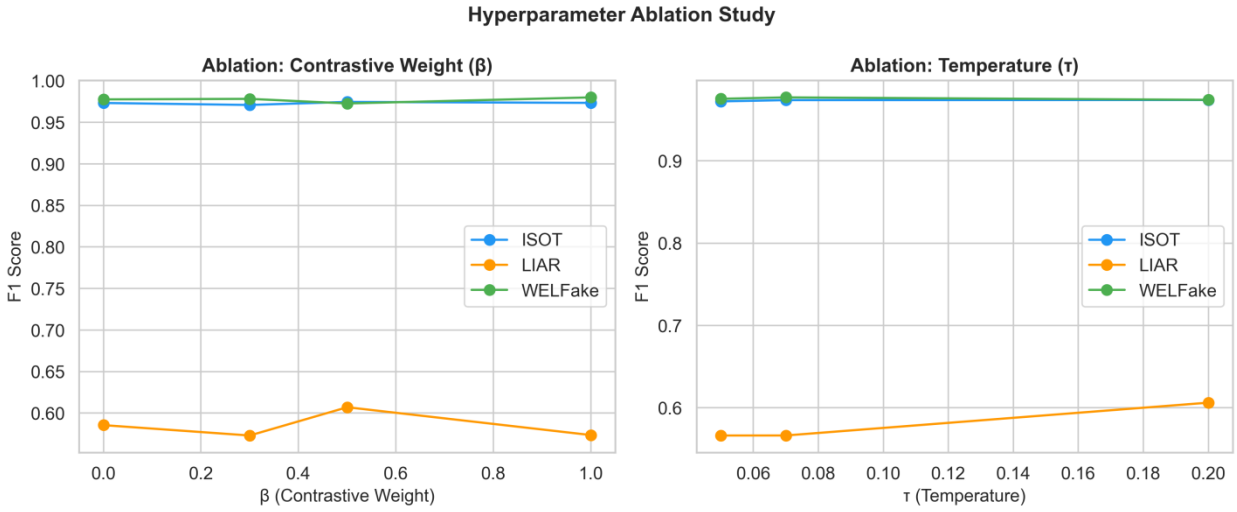
На підставі аналізу досягнутих результатів можна зробити висновки щодо параметрів абеляції. Методи ISOT і WELFake не чутливі до контрастних гіперпараметрів: за умови зміни цих параметрів показник F1 змінюється менше ніж на 0,7% для всіх налаштувань. LIAR демонструє значну чутливість – значення $\beta = 0,5$ забезпечує найкращий показник LIAR F1 (0,607), що на 2,2% вищий, ніж за умови об'єднання лише за CE $(\beta = 0,0,\ F1 = 0,585)$. Щодо температури, то $\tau = 0,2$ дає найкращий результат LIAR (0,606). Тенденції абляції зображено на рис. 3. Бачимо, що показники ISOT і WELFake залишаються стабільними, тоді як LIAR демонструє значну чутливість: найкращі результати досягаються за умов $\beta = 0,5$ і $\tau = 0,2$.

Простір вбудовування було проаналізовано візуально. На рис. 4 подано приклад унаочнення у форматі t-SNE вбудовувань [CLS] для базової моделі ISOT (ліворуч) і контрастивної моделі з об'єднаним пулом (праворуч).

На наведеному прикладі бачимо, що в разі застосування базової моделі, навченої тільки на ISOT, набори даних і класи перемішані, а за умови використання контрастивної об'єднаної моделі набір LIAR утворює окремий кластер, тоді як ISOT та WELFake демонструють покращене розділення реальних і фальшивих даних. Отже, базова модель створює безладний простір вбудовування, в якому зразки з усіх наборів даних і класів перемішані між собою. Це підтверджує, що модель, навчена на ISOT, не має корисної структури представлення для інших

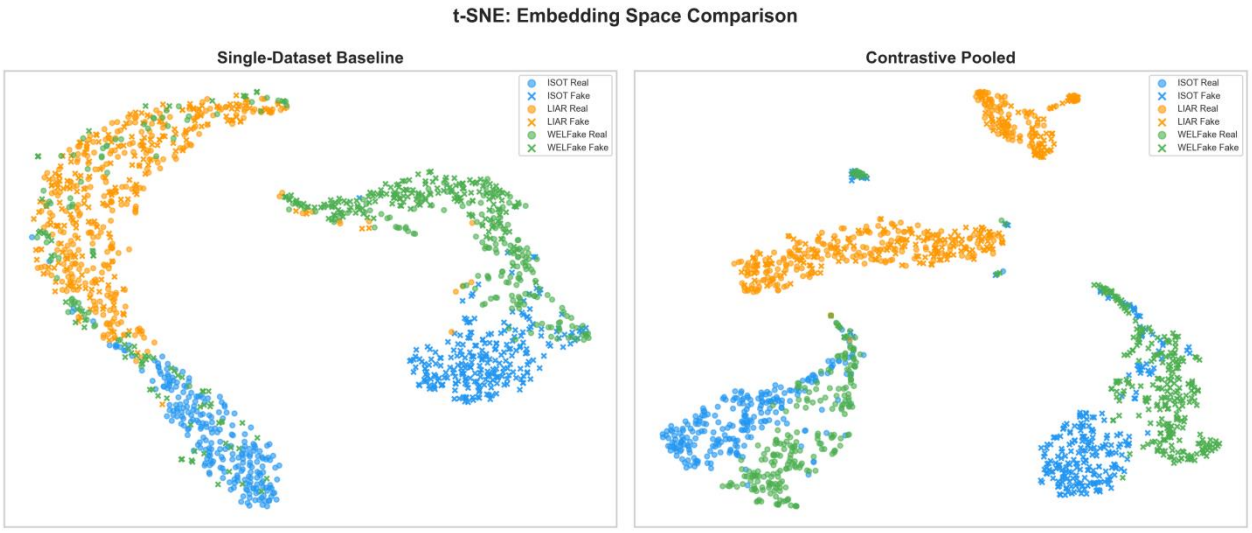
ISSN 3083-7715 (online) Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)
ISSN 3083-7650 (0135-1710) (print) Management Information System and Devices. 2026. No. 2 (189)

наборів даних. Контрастивна модель створює значно більш впорядкований простір: вислови з набору LIAR утворюють окремий кластер, що відображає їх різні текстові характеристики, тоді як ділянки ISOT і WELFake демонструють більш чітке розмежування між справжніми й фальшивими даними.



a b

Рис. 3. Абляція залежно від контрастного коефіцієнта β (*a*) і температури τ (*b*)



a b

Рис. 4. Візуалізація вбудовувань [CLS] за допомогою t-SNE (кола – справжні дані; хрестики – фальшиві дані):
a – базова модель, навчена виключно на наборі ISOT; *b* – контрастивна об'єднана модель

Обговорення результатів дослідження

Отримані результати надають кількісні докази того, що контрольні показники для окремих наборів даних створюють глибоко оманливу картину можливостей виявлення. Перехід від 100% F1 (внутрішньодоменний) до 0,5% F1 (міждоменний) для LIAR→WELFake є фундаментальним режимом відмови, а не просто зниження продуктивності. Тому оцінювання наборів даних має стати стандартною практикою в дослідженнях, присвячених виявленню фейкових новин. Ступінь невдачі перенесення навіть між двома наборами даних повного тексту (ISOT→WELFake: 0,239 F1) свідчить про те, що моделі навчаються розпізнавати джерело, а не виявляти обман. Статті *Reuters* мають певні стилістичні особливості, які слугують майже ідеальним показником "справжності" в ISOT, але ці особливості відсутні в джерелах, що не належать до *Reuters*.

Важливим висновком є те, що об'єднане навчання виключно на даних CE $(\beta = 0,0)$ вже

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

забезпечує більшу частину покращення результатів на інших наборах даних, тоді як контрастивне навчання надає додаткове покращення для складних доменів, як-от LIAR. Це свідчить про те, що доступ до різноманітних навчальних прикладів є основним чинником узагальнення, тоді як контрастивна задача допомагає вдосконалити репрезентації саме для тих доменів, де класифікаційний сигнал є найслабшим.

Результати дослідження дають змогу розширити підхід до корпоративної безпеки, де інформаційні ризики розглядаються нарівні з фінансовими чи операційними, а запропоновані інструменти для виявлення фейкових новин можуть стати невід'ємним складником стратегічного управління бізнесом. Сучасна галузева практика навчання моделей на єдиному наборі даних та їх впровадження, як це роблять соціальні мережі, агрегатори новин і системи модерації контенту, у цьому разі є принципово хибною. Запропонована матриця перенесення кількісно оцінює масштаб ризику: система, що демонструє 100% точність під час внутрішнього оцінювання, може працювати майже на рівні випадковості, коли стикається з новинами із джерел, не поданих у її навчальних даних.

Запропонований підхід контрастивного об'єднаного адаптаційного навчання відповідає сучасним потребам, створюючи єдину модель, яку можна впровадити в новинах різних сфер. Замість того, щоб підтримувати окремі моделі для різних типів контенту (великі статті, стислі політичні заяви, новини з різних джерел), організації можуть застосувати одну уніфіковану модель з високою ефективністю у всіх протестованих сферах. Це зменшує інженерну складність, витрати на інфраструктуру й ризик виникнення сліпих зон, пов'язаних із конкретною сферою, у процесах модерації контенту.

Методологія дедуплікації в запропонованому підході також має практичну цінність для організацій, що створюють навчальні набори даних із багатьох публічних джерел. Наприклад, виявлення 38 285 статей, що дублюються між двома широко використовуваними еталонами, демонструє, що забруднення наборів даних є реальною проблемою, яка може непомітно завищувати показники продуктивності.

Крім того, результати абляції дають практичні рекомендації щодо впровадження: контрастна вага $\beta = 0,5$ і температура $\tau = 0,2$ забезпечують оптимальну міждоменну генералізацію, а висновок про те, що лише об'єднане навчання дає змогу отримати більшу частину переваг, означає, що організації, які не мають ресурсів для створення інфраструктури контрастного навчання, все одно можуть досягти значного поліпшення генералізації за допомогою простого об'єднання даних із різних джерел.

Проведене дослідження обмежується наборами даних англійською мовою, оскільки багатомовна генералізація створює додаткові виклики. Дедуплікація WELFake скорочує набір даних до статей, які не належать до *Reuters*, що потенційно змінює рівень складності. Бінаризація LIAR зводить до нуля значущі градації, що потенційно може спричинити шум міток. Контрастивний підхід вимагає одночасного доступу до декількох наборів даних під час навчання, що не завжди є можливим в умовах обмеженої конфіденційності.

Висновки

У статті розглянуто систематичне дослідження перехресного узагальнення наборів даних у виявленні фейкових новин, продемонстровано, що оцінка одного набору даних створює принципово спотворену картину прогресу. Побудована матриця перенесення виявляє катастрофічну невдачу узагальнення: досягнення 100% F1 на базі DeBERTa-v3 падає до 0,5% F1 за умови перенесення на різні набори даних із середньою деградацією, що перевищує 75%. Для розв'язання цієї проблеми запропоновано контрастивне об'єднане адаптаційне навчання. Метод поєднує перехресну ентропійну класифікацію з контрольованим контрастивним навчанням одночасно на кількох наборах даних, що надає такі досягнення: 97,5% F1 на ISOT, 57,4% на LIAR та 98,4% на WELFake – із покращенням до 97,9% абсолютного F1 порівняно з базовими перенесеннями. Аналіз абляції підтверджує, що контрастна мета особливо корисна для більш складних доменів класифікації.

Наукова новизна. У роботі вперше застосовано контрольоване контрастивне навчання саме для міжнаборової адаптації доменів у виявленні фейкових новин на основі тексту. На відміну від попередніх контрастивних підходів щодо міжмодального вирівнювання між текстом і зображеннями, а також від діагностичних досліджень між наборами даних, які виявляють проблему узагальнення без пропозиції рішень, це дослідження поєднує кількісне оцінювання

ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

проблеми за допомогою контрольованої матриці перенесення $N \times N$ із конкретною стратегією навчання, що породжує єдину модель, ефективну в різних новинних доменах. Додатковим методологічним внеском є виявлення й вилучення статей, що дублюються між наборами даних, тобто усунення забруднення даних, яке має значення для всіх майбутніх досліджень із використанням цих бенчмарків.

Практична цінність запропонованого підходу виходить за межі академічних бенчмарків: інструменти виявлення фейкових новин, здатні до узагальнення між доменами, стають невід'ємною частиною стратегічного управління бізнесом і логічно вписуються в ширший контекст трансформацій сучасного інформаційного середовища.

Напрями подальшої роботи передбачають розширення фреймворку на додаткові бенчмарки й багатомовні набори даних, вивчення більш складних методів адаптації предметної галузі (змагальне навчання, метанавчання), дослідження взаємодії між упровадженням лінгвістичних ознак і порівняльним навчанням, а також вивчення того, як продуктивність масштабується залежно від кількості об'єднаних наборів даних.

Конфлікт інтересів

Автори заявляють, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

Фінансування

Це дослідження було проведено без фінансової підтримки.

Доступність даних

Дані будуть надані за обґрунтованим запитом. Набори даних ISOT, LIAR і WELFake, використані в цьому дослідженні, є загальнодоступними у відповідних оригінальних джерелах [27–29]. Дедуплікований варіант WELFake та експериментальний код будуть надані за обґрунтованим запитом.

Використання засобів штучного інтелекту

DeepL Translator (DeepL GmbH, нейромережева модель машинного перекладу на основі трансформерів) використовувався для створення попередніх перекладів розд. 1 та 3 з англійської на українську, а також для перекладу анотації з української на англійську. Весь перекладений текст автори перевірили вручну й відредагували для забезпечення термінологічної точності та стилістичної узгодженості з мовою перекладу.

References

1. Herasymov, S., Tkachov, A., Bazarnyi, S. (2024), "Complex method of determining the location of social network agents in the interests of information operations", *Advanced Information Systems*, Vol. 8, No. 1, pp. 31–36. DOI: https://doi.org/10.20998/2522-9052.2024.1.04
2. Podorozhniak, A., Liubchenko, N., Oliinyk, V., Roh, V. (2023), "Research application of the spam filtering and spammer detection algorithms on social media and messengers", *Advanced Information Systems*, Vol. 7, No. 3, pp. 60–66. DOI: https://doi.org/10.20998/2522-9052.2023.3.09
3. Allcott, H., Gentzkow, M. (2017), "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 211–236. DOI: https://doi.org/10.1257/jep.31.2.211
4. Bakir, V., McStay, A. (2018), "Fake News and The Economy of Emotions", *Digital Journalism*, Vol. 6, No. 2, pp. 154–175. DOI: https://doi.org/10.1080/21670811.2017.1345645
5. Datsenko, S. (2025), "Comparative Study of Machine Learning Approaches for Fake News Detection on Social Platforms", *2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek)*, pp. 1–5. DOI: https://doi.org/10.1109/KhPIWeek61436.2025.11288402
6. Christodoulou, C., Salamanos, N., Leonidou, P., Papadakis, M., Sirivianos, M. (2023), "Identifying misinformation on YouTube through transcript contextual analysis with Transformer models", *arXiv preprint*, arXiv: 2307.12155. DOI: https://doi.org/10.48550/arXiv.2307.12155
7. Datsenko, S. (2025), "Neural architecture comparison for fact verification on FEVER dataset", *Control, Navigation and Communication Systems*, Vol. 3, No. 81m pp. 72-75. DOI: https://doi.org/10.26906/SUNZ.2025.3.072

8. Shen, Y., Liu, Q., Guo, N., Yuan, J., Yang, Y. (2023), "Fake news detection on social networks: A survey", *Applied Sciences*, Vol. 13, No. 21, article 11877. DOI: https://doi.org/10.3390/app132111877
9. Capuano, N., Fenza, G., Loia, V., Nota, F. D. (2023), "Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review", *Neurocomputing*, Vol. 530, pp. 91–103. DOI: https://doi.org/10.1016/j.neucom.2023.02.005
10. Nickerson, R. S. (1998), "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises", *Review of General Psychology*, Vol. 2, No. 2, pp. 175-220. DOI: https://doi.org/10.1037/1089-2680.2.2.175.
11. Ozbay, F. A., Alatas, B. (2020), "Fake news detection within online social media using supervised artificial intelligence algorithms", *Physica A: Statistical Mechanics and Its Applications*, Vol. 540, article 123174. DOI: https://doi.org/10.1016/j.physa.2019.123174
12. Abualigah, L., Al-Ajlouni, Y. Y., Daoud, M. S., Altalhi, M., Migdady, H. (2024), "Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe", *Social Network Analysis and Mining*, Vol. 14, article 40. DOI: https://doi.org/10.1007/s13278-024-01198-w
13. Sastrawan, I. K., Bayupati, I. P. A., Arsa, D. M. S. (2021), "Detection of fake news using deep learning CNN-RNN based methods", *ICT Express*, Vol. 8, pp. 396–408. DOI: https://doi.org/10.1016/j.icte.2021.10.003
14. Nasir, J. A., Khan, O. S., Varlamis, I. (2021), "Fake news detection: A hybrid CNN-RNN based deep learning approach", *International Journal of Information Management Data Insights*, Vol. 1, No. 1, article 100007. DOI: https://doi.org/10.1016/j.jjimei.2020.100007
15. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of NAACL-HLT*, pp. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423
16. Liu, Y., Ott, M., Goyal, N. et al. (2019), "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *arXiv preprint*, arXiv: 1907.11692. DOI: https://doi.org/10.48550/arXiv.1907.11692
17. He, P., Gao, J., Chen, W. (2023), "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing", *Proceedings of ICLR.* DOI: https://doi.org/10.48550/arXiv.2111.09543
18. Hoy, N., Koulouri T. (2022), "Exploring the Generalisability of Fake News Detection Models", *Proceedings of IEEE International Conference on Big Data*, pp. 5731–5740. DOI: https://doi.org/10.1109/BigData55660.2022.10020583
19. Gruensteidl, M. N., Kirrane, S. (2026), "A Comparison and Critical Reflection of Information Disorder Detection Techniques: Performing a Cross-Data and Cross-Model Evaluation", *Information Fusion*, Vol. 127, Part B, article 103806. DOI: https://doi.org/10.1016/j.inffus.2025.103806
20. Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020), "A Simple Framework for Contrastive Learning of Visual Representations", *Proceedings of ICML*. DOI: https://doi.org/10.48550/arXiv.2002.05709
21. Khosla, P., Teterwak, P., Wang, C. et al. (2020), "Supervised Contrastive Learning", *Proceedings of NeurIPS*, Vol. 33, pp. 18661–18673. DOI: https://doi.org/10.48550/arXiv.2004.11362
22. Gao, T., Yao, X., Chen, D. (2021), "SimCSE: Simple Contrastive Learning of Sentence Embeddings", *Proceedings of EMNLP*, pp. 6894–6910. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.552
23. Gunel, S., Du, J., Conneau, A., Stoyanov, V. (2021), "Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning", *Proceedings of ICLR*. DOI: https://doi.org/10.48550/arXiv.2011.01403
24. Bhattacharjee, A., Kumarage, T., Moraffah, R., Liu H. (2023), "ConDA: Contrastive Domain Adaptation for AI-generated Text Detection", *Proceedings of IJCNLP-AACL*, pp. 598–610. DOI: https://doi.org/10.18653/v1/2023.ijcnlp-main.40
25. Yan, F., Zhang, M., Wei, B., Ren, K., Jiang, W. (2024), "SARD: Fake news detection based on CLIP contrastive learning and multimodal semantic alignment", *Journal of King Saud University – Computer and Information Sciences*, Vol. 36, article 102160. DOI: https://doi.org/10.1016/j.jksuci.2024.102160
26. Shen, X., Huang, M., Hu, Z., Cai, S., Zhou, T. (2024), "Multimodal Fake News Detection with Contrastive Learning and Optimal Transport", Frontiers in Computer Science, Vol. 6, article 1473457. DOI: https://doi.org/10.3389/fcomp.2024.1473457
27. ISOT Research Lab (2024), "ISOT Fake News Dataset", University of Victoria, available at: https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/ (last accessed 20.04.2026).
28. Wang, W. Y. (2017), "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection", *Proceedings of ACL*, pp. 422–426. DOI: https://doi.org/10.18653/v1/P17-2067
29. Verma, P. K., Agrawal, P., Amorim, I., Prodan, R. (2021), "WELFake: Word Embedding Over Linguistic Features for Fake News Detection", *IEEE Transactions on Computational Social Systems, Vol*. 8, No. 4, pp. 881–893. DOI: https://doi.org/10.1109/TCSS.2021.3068519

Received (Надійшла) 22.03.2026
Accepted for publication (Прийнята до друку) 16.04.2026
Publication date (Дата публікації) 29.05.2026

ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

Відомості про авторів / About the Authors

Даценко Сергій Сергійович – Національний технічний університет "Харківський політехнічний інститут", аспірант кафедри комп'ютерної інженерії та програмування, Харків, Україна;
Serhii Datsenko – National Technical University "Kharkiv Polytechnic Institute", Postgraduate Student of the Computer Engineering and Programming Department, Kharkiv, Ukraine;
e-mail: sergdacenko@gmail.com
ORCID ID: https://orcid.org/0000-0001-9514-0433

Кучук Георгій Анатолійович – доктор технічних наук, професор, Національний технічний університет "Харківський політехнічний інститут", професор кафедри комп'ютерної інженерії та програмування, Харків, Україна;
Heorhii Kuchuk – Doctor of Technical Sciences, Professor, National Technical University "Kharkiv Polytechnic Institute", Professor of Computer Engineering and Programming Department, Kharkiv, Ukraine;
e-mail: kuchuk56@ukr.net
ORCID ID: http://orcid.org/0000-0002-2862-438X

CROSS-DATASET GENERALIZATION FOR FAKE NEWS DETECTION USING CONTRASTIVE FINE-TUNING

The needs of modern industrial business are increasingly challenged by the spread of fake news, which can significantly affect the reputation, customer trust, and financial performance of industrial companies. Existing detection systems demonstrate nearly perfect accuracy on individual benchmark datasets; however, they suffer catastrophic failure when applied to information from different sources, raising serious concerns about their readiness for real-world deployment. The subject of the study is methods for fake news detection in structurally diverse datasets. The aim of the study is to develop and validate an approach for fake news detection in datasets with structurally diverse content using contrastive fine-tuning, enabling the creation of a unified transformer-based model. The objectives of the study include: establishing baseline performance indicators for individual datasets using the DeBERTa-v3-base language model on three different fake news datasets; constructing a systematic cross-dataset transfer matrix to quantitatively evaluate generalization failures; proposing and evaluating contrastive joint fine-tuning as a domain adaptation approach; and conducting ablation studies of contrastive hyperparameters. **Research method.** The DeBERTa-v3-base language model is fine-tuned on three fake news datasets: ISOT, LIAR, and WELFake. The combined training objective integrates cross-entropy classification with supervised contrastive loss on merged datasets using balanced sampling across different datasets. **Results.** The cross-dataset transfer matrix demonstrates catastrophic generalization failure: a model achieving 100% F1 on ISOT attains only 23.9% on WELFake and 40.3% on LIAR. The proposed contrastive joint model produces a unified model that simultaneously achieves 97.5% F1 on ISOT, 57.4% on LIAR, and 98.4% on WELFake, with an absolute F1 improvement of up to 97.9% compared to baseline transfers. **Conclusions.** Ablation analysis confirms that the contrastive objective is particularly beneficial in more challenging domains. Contrastive joint fine-tuning enables the development of models suitable for deployment in multi-source news environments.

Keywords: fake news detection; cross-dataset generalization; transformer language model; transfer learning; natural language processing.

Бібліографічні описи / Bibliographic descriptions

Даценко С. С., Кучук Г. А. Міжвибіркова узагальненість для виявлення фейкових новин із використанням контрастивного адаптаційного навчання. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 153–164. DOI: https://doi.org/10.30837/0135-1710.2026.189.153

Datsenko, S., Kuchuk, H. (2026), "Cross-Dataset Generalization for Fake News Detection Using Contrastive Fine-Tuning", *Management Information System and Devices*, No. 2 (189), P. 153–164. DOI: https://doi.org/10.30837/0135-1710.2026.189.153