.МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Навчально-науковий інститут інноваційних освітніх технолллогій
Кафедра інформаційно-обчислювальних систем і управління

ЮРКІВ Христина Володимирівна

Нейромережевий метод розпізнавання діпфейків на основі характеристик обличчя / Neural network method for deepfake detection based on facial features

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп’ютерні науки

Кваліфікаційна робота

Виконав студент групи КНмз-21
Х. В. Юрків
_____

Науковий керівник:
д.т.н., доцент Х.В. Ліп’яніна-Гончаренко
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 20___ р.
В.о. завідувача кафедри
_____ Н.В. Дзюбановська

ТЕРНОПІЛЬ – 2025

Навчально-науковий інститут інноваційних освітніх
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «магістр»
спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ Н.М. Васильків
«____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
ЮРКІВ Христина Володимирівна

(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи

Нейромережевий метод розпізнавання діпфейків на основі характеристик обличчя / Neural network method for deepfake detection based on facial features

керівник роботи к.т.н., доцент Х. В. Ліп'яніна-Гончаренко
затверджені наказом по університету від 20 грудня 2024 року № 938.

2. Строк подання студентом закінченої кваліфікаційної роботи 1 грудня 2025 р.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити

- - провести аналіз сучасних методів і технологій виявлення діпфейків, з особливим акцентом на підходах, орієнтованих на аналіз facial features та просторово-часової динаміки обличчя;
- визначити найбільш інформативні ознаки обличчя для задачі ідентифікації фейкових відео та підходи до їхньої формалізації;
- спроєктувати архітектуру програмного модуля, який забезпечує детекцію обличчя у відеокадрах, виділення ключових точок (facial landmarks) та формування ознакового простору для подальшої класифікації;
- реалізувати програмний модуль, що здійснює обробку відео, екстракцію характеристик обличчя та класифікацію відеофрагментів за ознакою достовірності/підробки;
- провести експериментальні дослідження на відкритих наборах даних реальних і синтетичних відео (зокрема Deepfake Detection Challenge) з оцінюванням точності, повноти, F1-міри та інших показників якості класифікації;
- виконати порівняльний аналіз нейромережевого підходу із евристичним критерієм, що ґрунтується на аномаліях у структурі та поведінці обличчя, і оцінити можливості їхньої комбінованої інтеграції в єдину систему розпізнавання діпфейків.

5. Перелік графічного матеріалу у роботі

- трирівнева архітектура методу розпізнавання діпфейків на основі аналізу обличчя;
- структурна схема програмного модуля ідентифікації ознак облич у діпфейках

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 20 грудня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2025 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2025 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 20.05.2025 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 28.10. 2025 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 11.11.2025 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 25.11.2025 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту.	до 1.12.2025 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 04.12.2025 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.12. 2025 р.	

Студент _____ Х.В. Юрків
підпис

Керівник роботи _____ к.т.н., доцент Х.В. Ліп’яніна-Гончаренко
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему «Нейромережевий метод розпізнавання діпфейків на основі характеристик обличчя» на здобуття освітнього ступеня «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 131 сторінки і містить 8 ілюстрацій, 2 таблиці, 2 додатки та 37 використаних джерел.

Метою даної кваліфікаційної роботи є розроблення інтелектуального модуля ідентифікації ознак облич у діпфейкових відео на основі методів комп'ютерного зору та машинного навчання для підвищення достовірності виявлення штучно зміненого відеоконтенту.

Методи досліджень: аналіз і синтез наукових джерел, системний підхід, алгоритми комп'ютерного зору для виявлення та трекінгу обличчя, методи виділення ознак обличчя, логістична регресія та інші методи машинного навчання, експериментальне моделювання на наборах даних діпфейків.

Результати дослідження: розроблено архітектуру програмного модуля, сформовано простір ознак обличчя та запропоновано алгоритм класифікації відео як REAL/FAKE на основі поєднання евристичних правил і логістичної регресії. Проведені експерименти показали підвищення показників точності та F1-міри порівняно з базовими евристичними підходами.

Результати роботи можуть бути використані під час створення систем автоматизованої модерації відеоконтенту, сервісів моніторингу діпфейків та в освітньому процесі при вивченні дисциплін із комп'ютерного зору, штучного інтелекту та кібербезпеки.

Ключові слова: ДІПФЕЙК, ВИЯВЛЕННЯ ДІПФЕЙКІВ, АНАЛІЗ ОБЛИЧЧЯ, ОЗНАКИ ОБЛИЧЧЯ, ЛОГІСТИЧНА РЕГРЕСІЯ, ГЛИБИННЕ НАВЧАННЯ, ВІДЕОКОНТЕНТ, ІНФОРМАЦІЙНА БЕЗПЕКА.

ABSTRACT

Qualification work on the topic «Neural network method for deepfake detection based on facial features » for Master's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 131 pages and it contains 8 figures, 2 table, 2 annexes and 37 sources.

The aim of this qualification work is to develop an intelligent module for the identification of facial features in deepfake videos based on computer vision and machine learning methods in order to increase the reliability of detecting artificially manipulated video content.

Research methods: analysis and synthesis of scientific sources, a systems approach, computer vision algorithms for face detection and tracking, methods for extracting facial features, logistic regression and other machine learning methods, as well as experimental modelling on deepfake datasets.

Research results: the architecture of the software module has been developed, the facial feature space has been constructed, and an algorithm for classifying videos as REAL/FAKE based on a combination of heuristic rules and logistic regression has been proposed. The experiments conducted have shown an improvement in accuracy and F1-score compared to basic heuristic approaches.

The results of this work can be used in the development of systems for automated video content moderation, deepfake monitoring services, and in the educational process when teaching courses in computer vision, artificial intelligence, and cybersecurity.Keywords: DEEPFAKE, DEEPFAKE DETECTION, FACE ANALYSIS, FACIAL FEATURES, LOGISTIC REGRESSION, DEEP LEARNING, VIDEO CONTENT, INFORMATION SECURITY.

ЗМІСТ

ВСТУП

Сучасний етап розвитку інформаційного суспільства характеризується стрімким зростанням обсягів мультимедійного контенту та широким застосуванням генеративних моделей штучного інтелекту для обробки зображень і відео. Одним із найнебезпечніших проявів таких технологій стали діпфейки — синтетичні відео та зображення, у яких зовнішність або мова людини змінюються за допомогою глибинних нейронних мереж при збереженні високого рівня візуальної та аудіальної правдоподібності. У результаті виникає якісно новий клас загроз інформаційній безпеці, пов’язаних із можливістю масштабної маніпуляції суспільною думкою, дискредитації публічних осіб, шантажу, шахрайства та підриву довіри до цифрових медіа.

Актуальність теми кваліфікаційної роботи зумовлена, по-перше, стрімким вдосконаленням генеративних архітектур (GAN, StyleGAN, diffusion-моделі), які дають змогу створювати діпфейки, практично нерозрізнені на «на око» навіть для підготовлених фахівців. По-друге, зростає кількість прикладних сценаріїв, у яких такі синтетичні відео використовуються як інструмент дезінформаційних кампаній, кібератак та соціальної інженерії, зокрема в політичній, фінансовій та правовій сферах. По-третє, на тлі воєнної агресії проти України та посилення гібридних інформаційних операцій особливого значення набувають надійні інструменти автоматизованої перевірки достовірності відеоконтенту, у тому числі з урахуванням обмежених ресурсів обчислювальної інфраструктури.

Традиційні підходи до виявлення підробленого відео, що базуються на аналізі метаданих, статистичних характеристик зображень або простих евристичних ознак, виявилися недостатньо ефективними в умовах сучасних генеративних моделей. Вони часто не забезпечують прийнятної якості класифікації в умовах зміни домену (інший набір даних, інший тип діпфейка, інший рівень компресії відео) або вимагають значних обчислювальних витрат.

На цьому тлі перспективним напрямом є методи, орієнтовані на аналіз формалізованих ознак обличчя — facial landmarks, мікровиразів, геометричних та симетричних характеристик, які відображають фізіологічні й поведінкові особливості людини і суттєво складніші для достовірного відтворення генеративними моделями.

Метою кваліфікаційної роботи є розроблення нейромережевого методу розпізнавання діпфейків на основі формалізованих ознак обличчя та створення програмного модуля ідентифікації таких ознак у відеоданих із використанням сучасних методів комп'ютерного зору та глибинного навчання.

Для досягнення поставленої мети в роботі необхідно розв'язати такі основні завдання:

1) провести аналіз сучасних методів і технологій виявлення діпфейків, з особливим акцентом на підходах, орієнтованих на аналіз facial features та просторово-часової динаміки обличчя;

2) визначити найбільш інформативні ознаки обличчя для задачі ідентифікації фейкових відео та підходи до їхньої формалізації;

3) спроєктувати архітектуру програмного модуля, який забезпечує детекцію обличчя у відеокадрах, виділення ключових точок (facial landmarks) та формування ознакового простору для подальшої класифікації;

4) реалізувати програмний модуль, що здійснює обробку відео, екстракцію характеристик обличчя та класифікацію відеофрагментів за ознакою достовірності/підробки;

5) провести експериментальні дослідження на відкритих наборах даних реальних і синтетичних відео (зокрема Deepfake Detection Challenge) з оцінюванням точності, повноти, F1-міри та інших показників якості класифікації;

6) виконати порівняльний аналіз нейромережевого підходу із евристичним критерієм, що ґрунтується на аномаліях у структурі та поведінці обличчя, і оцінити можливості їхньої комбінованої інтеграції в єдину систему розпізнавання діпфейків.

Об’єктом дослідження в даній кваліфікаційній роботі є процес виявлення діпфейків у відеоданих у задачах забезпечення інформаційної безпеки.

Предметом дослідження є нейромережеві методи розпізнавання діпфейків, що базуються на виділенні, формалізації та аналізі характеристик обличчя у відеопослідовностях, зокрема facial landmarks, мікровиразів та пов’язаних з ними просторово-часових залежностей.

Методи дослідження. Для розв’язання поставлених завдань у роботі використовуються методи аналізу та синтезу, системний підхід до моделювання процесів виявлення діпфейків, методи статистичного аналізу та візуалізації даних, методи машинного навчання та глибинних нейронних мереж, алгоритми комп’ютерного зору для детекції обличчя та facial landmarks, а також експериментальне моделювання на основі відкритих відеодатасетів. При програмній реалізації застосовуються сучасні бібліотеки й фреймворки глибинного навчання та обробки зображень, що забезпечують відтворюваність експериментів і можливість подальшої інтеграції розробленого модуля в прикладні системи.

Практичне значення одержаних результатів полягає в розробленні програмного модуля, який може бути використаний як окремий інструмент для попереднього скринінгу відеоконтенту на наявність діпфейків, а також як складова частина комплексних систем моніторингу інформаційних загроз, платформ модерації користувацького контенту або судово-комп’ютерної експертизи. Запропонований підхід може бути адаптований для інтеграції у корпоративні рішення, сервіси перевірки достовірності цифрових доказів та системи підтримки прийняття рішень у сфері кібербезпеки. Крім того, результати роботи можуть бути використані в освітньому процесі при викладанні дисциплін, пов’язаних із комп’ютерним зором, машинним навчанням та інформаційною безпекою.

Наукова новизна одержаних результатів полягає в удосконаленні підходу до виявлення діпфейків шляхом комплексного аналізу

формалізованих ознак обличчя (геометрії, структури facial landmarks і мікродинаміки міміки) та розробленні трирівневої архітектури програмного модуля з інтеграцією евристичних критеріїв аномальності й нейромережевого класифікатора, що підвищує стійкість розпізнавання до різних типів синтетичного відеоконтенту.

Апробація результатів дослідження. Основні теоретичні положення та практичні результати роботи були апробовані на провідних міжнародних наукових заходах, зокрема: 5th IEEE KhPI Week on Advanced Technology (KhPIWeek 2024, Харків, Україна, жовтень 2024 р.); 1st International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development (AISD 2024, Тернопіль, Україна, 10–11 травня 2024 р.); 7th International Workshop on Computer Modeling and Intelligent Systems (CMIS 2024, Запоріжжя, Україна, 3 травня 2024 р.). Також результати дослідження відображено в статтях: Lipianina-Honcharenko, K., Komar, M., Bohuta, H., Ihnatiev, I., Yurkiv, K., Illiashenko, O., & Bilovus, L. (2025). A general method for real-time detection of information threats with a Ukraine case study. Radioelectronic and Computer Systems, 2025(3), 202–230 (журнал належить до квартилю Q2 за базою Scopus); Lipianina-Honcharenko, K., Ihnatiev, I., Bohuta, H., Yurkiv, K., & Novosad, S. (2025). Rule-based method for aligning media content with Ukrainian legislation. Матеріали доповідей опубліковано у збірниках праць, проіндексованих у наукометричній базі Scopus, що підтверджує наукову й практичну значущість одержаних результатів. Копії поблікацій розміщені в додатку Б.

1 ТЕОРЕТИЧНІ ЗАСАДИ ТА СУЧАСНІ ПІДХОДИ ДО ІДЕНТИФІКАЦІЇ ДІПФЕЙКІВ

1.1 Загрози, пов’язані з діпфейками, та їхній вплив на інформаційну безпеку

Технологія діпфейків, що базується на генеративних моделях (GAN, автоенкодери та інші архітектури глибинного навчання), стала одним із найбільш дискусійних явищ сучасної цифрової епохи. Її поширення створює багаторівневі ризики — від індивідуальної приватності до глобальної політичної стабільності. Проблема полягає не лише у високій реалістичності синтетичних матеріалів, але й у швидкій адаптації алгоритмів генерації до методів детекції, що формує своєрідну «гонку озброєнь» між творцями та дослідниками.

Діпфейки можна розглядати як інструмент подвійного призначення: з одного боку, вони мають потенціал для творчих і освітніх застосувань (наприклад, у кіноіндустрії чи віртуальній реальності), а з іншого — стають потужним засобом маніпуляції та дезінформації. Саме другий аспект становить найбільшу загрозу для інформаційної безпеки.

У контексті інформаційної безпеки діпфейки слід розглядати як багатовимірну загрозу, що охоплює політичну, економічну, соціальну та правову сфери. Їхній вплив не обмежується окремими інцидентами — він формує нову реальність цифрової комунікації, де автентичність інформації постійно ставиться під сумнів (таблиця 1.1).

- Політичний вимір. Синтетичні відео та аудіо, створені з використанням діпфейк-технологій, здатні радикально змінювати хід політичних процесів. Імітація заяв політичних лідерів, створення компрометуючих матеріалів або маніпулятивних звернень може впливати на електоральну поведінку, формувати викривлену громадську думку та слугувати інструментом інформаційних операцій у міжнародному середовищі.

- Економічний вимір. У корпоративному секторі діпфейки відкривають нові вектори атак — від шахрайських відеозвернень від імені керівництва до підроблених інструкцій, що вводять в оману співробітників. Особливо небезпечними є сценарії, пов’язані з обходом біометричних систем аутентифікації, що ставить під загрозу конфіденційність і фінансову безпеку організацій.

- Соціальний вимір. Масове поширення діпфейків спричиняє ерозію довіри до цифрових джерел інформації. У суспільстві формується атмосфера інформаційної невизначеності, де навіть достовірні матеріали можуть бути сприйняті як фальшиві. Це породжує феномен «постправди» та ускладнює критичне мислення в умовах медіаперевантаження.

- Правовий вимір. Юридична система стикається з викликами, пов’язаними з автентичністю цифрових доказів. Діпфейки підривають довіру до відео- та аудіоматеріалів, які традиційно вважалися надійними доказами у судовому процесі. Це вимагає перегляду нормативно-правових підходів до оцінки цифрового контенту та впровадження нових стандартів цифрової експертизи.

Таблиця 1.1 - Основні загрози діпфейків та їхній вплив

№	Тип загрози	Сфера впливу	Приклади наслідків	Рівень ризику
1.	Політична дезінформація	Суспільна довіра, виборчі процеси	Фальшиві виступи політиків, маніпуляція громадською думкою	Високий
2.	Кіберзлочинність і соціальна інженерія	Бізнес, кібербезпека	Імітація керівництва компаній, обходи біометрії, таргетований фішинг	Високий
3.	Репутаційні атаки	Індивідуальна та корпоративна репутація	Дискредитація публічних осіб, компрометація брендів	Середній–високий
4.	Порушення приватності	Особисті права	Порнографічний контент без згоди, психологічні травми	Високий
5.	Підрив довіри до медіа	Інформаційний простір	«Парадокс правди» — автентичні матеріали відкидаються як фейк	Високий

6.	Юридичні проблеми	Судова практика, правоохоронні органи	Недовіра до відео- та аудіодоказів	Середній

Діпфейки як феномен цифрової епохи становлять багатовекторну загрозу, що охоплює ключові сфери інформаційної безпеки — політичну, економічну, соціальну та правову. Їхня здатність імітувати реальність з високим ступенем достовірності підриває фундаментальні механізми довіри, автентифікації та правозастосування.

Аналіз показує, що діпфейки не лише ускладнюють верифікацію інформації, але й створюють нові сценарії зловживань — від маніпуляції громадською думкою до обходу біометричних систем. В умовах зростаючої доступності інструментів генерації синтетичного контенту, детекція діпфейків перетворюється на стратегічне завдання, яке потребує міждисциплінарного підходу, поєднання технічних, правових та етичних рішень.

1.2 Сучасні методи та технології виявлення діпфейків

Сучасні підходи до виявлення діпфейків охоплюють широкий спектр методів — від класичних згорткових нейронних мереж до трансформерів, оптичного потоку та аномалійно-орієнтованих моделей. Більшість рішень базуються на аналізі обличчя як ключового джерела артефактів, проте дедалі більшої уваги набувають просторово-часові моделі, здатні виявляти неузгодженості в динаміці виразів обличчя, мікрорухах та контексті сцени. У цьому підрозділі розглянуто низку репрезентативних досліджень, які демонструють кількісні результати на відомих наборах даних та задають орієнтири для подальшого розвитку методів, у тому числі нейромережевих рішень на основі характеристик обличчя.

Одним із базових орієнтирів для оцінки алгоритмів є набір даних FaceForensics++, де в роботі Rössler та співавторів було досліджено детектор на основі архітектури XceptionNet для виявлення маніпульованих облич на

різних типах підробок (DeepFakes, Face2Face, FaceSwap, NeuralTextures) [1]. Для зображень без втрат або з легкою компресією детектор досягає площі під ROC-кривою (AUC) на рівні 0,997, тоді як для сильно стиснутих відео AUC знижується до приблизно 0,955, але все ще істотно перевищує показники людини-експерта. Ці результати продемонстрували, що глибокі згорткові мережі, спеціально адаптовані до текстурних артефактів обличчя, можуть забезпечувати майже «ідеальну» точність на контрольованих наборах даних, проте чутливі до агресивної компресії.

Компактний підхід MesoNet (Meso-4 та MesoInception-4) спрямований на зменшення обчислювальних витрат при збереженні прийнятної точності виявлення відеопідробок [2]. У роботі Afchar та співавторів показано, що MesoNet досягає точності понад 98 % для виявлення класичних Deepfake-відео та близько 95 % для Face2Face-маніпуляцій на відповідних наборах даних, залишаючись придатним для майже реального часу. Завдяки невеликій глибині мережі та обмеженій кількості параметрів MesoNet часто використовується як базовий легковаговий детектор та як елемент більш складних ансамблевих або каскадних систем.

Цікавим напрямом є методи, що враховують неузгодженість між обличчям і контекстом сцени. У роботі Nirkin та співавторів запропоновано окремо аналізувати ознаки, виділені з області обличчя, та ознаки фонового контексту, а далі зіставляти їх у єдиному класифікаторі [3]. На наборі FaceForensics++ запропонований метод досягає AUC, порівнянної з найкращими на той час системами (понад 99 % для деяких піднаборів), а на Celeb-DF-v2 покращує AUC приблизно на 1,4 відсоткового пункту порівняно з XceptionNet-базою (з ~64,6 % до ~66,0 %). Таким чином демонструється, що врахування логічної відповідності між обличчям та оточенням знижує кількість хибнопозитивних спрацьовувань.

У роботі Janūtėnas та співавторів запропоновано сім різних моделей глибокого навчання для виявлення підроблених зображень, серед яких модель LRNet показала найвищі результати [4]. На наборі FaceForensics++ без

компресії LRNet досягає 99,7 % точності та 99,9 % AUC, а для сильно стиснутих зображень AUC залишається вище 91 %, що свідчить про відносну стійкість моделі до артефактів стиснення. Дослідження також демонструє, що поєднання класичних CNN-шарів із механізмами регуляризації та оптимізованими функціями втрат дозволяє зберігати високу якість класифікації за різних умов кодування відео.

Окремий клас підходів базується на аналізі оптичного потоку. У роботі Cozzolino та колег запропоновано CNN, що працює не з самими кадрами, а з полем оптичного потоку, оціненим між ними, з метою виявлення неузгодженостей у русі [5]. Показано, що така модель досягає точності понад 90 % у задачі виявлення діпфейків на кількох варіантах FaceForensics++, причому навіть для маніпуляцій, які не були присутні у навчальній вибірці, точність лишається вищою за 90 %. Це свідчить про здатність оптичного потоку фіксувати фундаментальні порушення природної динаміки обличчя, характерні для синтетичних відео.

Подальші дослідження у цьому напрямку фокусуються на удосконаленні оцінки оптичного потоку. Так, Nassif і Shahin запропонували метод, що поєднує вдосконалене оцінювання оптичного потоку з CNN-класифікатором, орієнтованим на виявлення діпфейків у складних умовах [6]. На експериментальних наборах даних модель досягає точності близько 82 %, перевищуючи базові конфігурації оптичного потоку без додаткових корекцій. Попри дещо нижчу абсолютну точність порівняно з найкращими зображеневими моделями, такі рішення є важливими для сценаріїв, де критичною є чутливість до дрібних динамічних артефактів.

Ще один перспективний напрям — використання функцій міміки та Facial Action Units (FAU). У роботі Tan та співавторів запропоновано метод FADE (Facial Action Dependencies Estimation), який моделює залежності між елементарними діями м’язів обличчя та зіставляє їх із природними патернами виразів [7]. Згідно з кількісними результатами, інтеграція FAU дає приріст AUC до 4,47 відсоткового пункту та помітно підвищує точність на основних

відеонаборах (FaceForensics++, Celeb-DF, DFDC) порівняно з базовими CNN-моделями без урахування динаміки міміки. Таким чином підтверджується гіпотеза, що так звані «порушення природної міміки» є інформативною ознакою для виявлення синтетичного відеоконтенту.

Для повноцінного врахування просторово-часових залежностей були розроблені гібридні моделі на кшталт CNN–LSTM–Transformer. Petmezas та співавтори запропонували модель, яка спочатку екстрагує просторові ознаки з обличчя за допомогою ResNet, потім аналізує послідовність кадрів через LSTM та завершує обробку трансформерним блоком для моделювання довготривалих залежностей [8]. На валідаційному наборі автори отримали AUC 90,82 % та F1-міру близько 88 %, а на наборі DeepFake Detection (DFD) AUC сягнув 97 % при використанні оптимальної комбінації фреймів. Отримані результати демонструють, що адаптивне поєднання згорткових і послідовнісних блоків дозволяє покращити баланс між точністю та узагальнювальною здатністю моделі.

Проблему узагальнення на нові типи діпфейків розглянуто в роботі Wang та співавторів, де запропоновано підхід до аномалійно-орієнтованого виявлення з використанням спеціальної стратегії «boundary-blurring» в ознаковому просторі [9]. На наборі FaceForensics++ модель досягає AUC 99,56 %, тоді як середнє AUC на п’яти зовнішніх наборах даних (у режимі крос-доменного тестування) становить 83,32 %, що на 6–19 відсоткових пунктів краще за попередні методи-аналогі. Таким чином, акцент робиться не лише на точності в межах однієї доменної вибірки, а й на стійкості моделі до змін у типах генеративних моделей, якості відео та умов зйомки.

В роботі Zhao та співавторів запропоновано ISTVT (Interpretable Spatial-Temporal Video Transformer) — інтерпретований просторово-часовий відео-трансформер, здатний одночасно аналізувати просторові артефакти на обличчі та часову неузгодженість між кадрами [10]. Архітектура використовує декомпозитну просторово-часову самоувагу та спеціальний «self-subtract» механізм для підсилення відмінностей між реальними та синтетичними

фрагментами; додатково запропоновано алгоритм релеванс-візуалізації, який показує важливі області в кадрі та часовій осі. Експерименти на великих наборах FaceForensics++, FaceShifter, DeeperForensics, Celeb-DF та DFDC демонструють сильні показники як для внутрішньодоменних, так і для крос-доменних сценаріїв, що підтверджує ефективність трансформерних архітектур для задачі виявлення діпфейків.

Як показує порівняльний аналіз сучасних методів і моделей детекції діпфейків (таблиця 1.2), найвищі значення AUC (до 0,99 і вище) досягаються глибокими згортковими моделями, зокрема XceptionNet та LRNet, але їхня ефективність помітно знижується за умов сильної компресії відео [1, 4]. Легковагові архітектури, такі як MesoNet, забезпечують дещо нижчу, але все ще високий рівень точності за значно менших обчислювальних витрат, що робить їх придатними для застосувань у режимі, близькому до реального часу [2]. Методи, що враховують контекст сцени, міміку та Facial Action Units, а також просторово-часову динаміку (оптичний потік, гібридні CNN–LSTM–Transformer, відео-трансформери), демонструють кращу стійкість до нових типів діпфейків і покращують AUC на кілька відсоткових пунктів [3, 5–8, 10]. Окрему групу становлять аномалійно-орієнтовані підходи, які хоч і поступаються за максимальною точністю на окремих наборах, проте забезпечують ліпше узагальнення на зовнішні датасети [9].

Виявлення ознак на обличчі залишається одним із найефективніших підходів у протидії діпфейковим технологіям, оскільки підробка зовнішності людини супроводжується порушенням природних геометричних, симетричних та поведінкових характеристик. Попри стрімкий розвиток генеративних моделей, саме аналіз реальних фізіологічних особливостей — положення очей, пропорцій обличчя, синхронності міміки та стабільності ключових точок — дає змогу виявляти приховані аномалії, які важко відтворити алгоритмами штучного інтелекту. На відміну від методів, що працюють лише з текстурами або глобальними характеристиками зображення, підхід із використанням ознак обличчя дозволяє оцінювати глибинні

властивості поведінки людини у кадрі, що підвищує точність класифікації навіть у випадку високоякісних діпфейків. Саме тому методи, орієнтовані на детекцію та аналіз ключових компонентів обличчя, залишаються актуальними.

Таблиця 1.2 - Порівняльний аналіз досліджень

№	Основна ознака	Метод / Архітектура	Точність / AUC (приблизно)	Джерело
1	Текстурні артефакти обличчя, чутливість до компресії	XceptionNet на FaceForensics++	AUC ≈ 0,997 (low-compression), AUC ≈ 0,955 (strong-compression)	[1]
2	Компактність моделі та робота майже в реальному часі	MesoNet (Meso-4, MesoInception-4)	Точність > 98 % (Deepfake), ≈ 95 % (Face2Face)	[2]
3	Узгодженість обличчя та контексту сцени	CNN з окремими гілками для обличчя та фону	AUC > 99 % (FF++), приріст ≈ +1,4 п.п. AUC на Celeb-DF-v2	[3]
4	Стійкість до компресії, детекція фальсифікації зображень	LRNet (глибока CNN)	Точність 99,7 %, AUC 99,9 % (без компресії), AUC > 91 % (strong-compr.)	[4]
5	Аналіз просторово-часової динаміки через оптичний потік	CNN по полю оптичного потоку	Точність > 90 % на кількох варіантах FaceForensics++	[5]
6	Покращене оцінювання оптичного потоку для детекції діпфейків	Оптичний потік + CNN-класифікатор	Точність ≈ 82 %	[6]
7	Міміка та Facial Action Units, порушення природних виразів обличчя	FADE (моделювання залежностей між FAU)	Приріст AUC до ≈ +4,47 п.п. на FF++, Celeb-DF, DFDC	[7]
8	Спільний аналіз простору й часу через гібридну архітектуру	CNN–LSTM–Transformer	AUC ≈ 90,82 % (val), F1 ≈ 88 %, AUC ≈ 97 % (DFD)	[8]
9	Узагальнення на нові типи діпфейків (аномалійно-орієнтований підхід)	Anomaly-based детектор з «boundary-blurring»	AUC 99,56 % (FF++), середнє AUC ≈ 83,32 % (5 зовнішніх датасетів)	[9]

10	Інтерпретований просторово-часовий аналіз відео через трансформер	ISTVT – Spatial-Temporal Video Transformer	State-of-the-art / AUC > 0,90 на FF++, Celeb-DF, DFDC та ін.	[10]

1.3 Постановка задачі та завдання дослідження

Актуальність дослідження зумовлюється тим, що цифровий відеоконтент став одним із ключових носіїв суспільно значущої інформації, а його споживання відбувається у високошвидкісних каналах комунікації (соціальні мережі, месенджери, стримінгові сервіси), де перевірка достовірності часто відсутня або має реактивний характер. У таких умовах довіра до відео як «доказу» дедалі частіше визначає репутаційні, фінансові, правові та безпекові наслідки, що робить критично важливим розвиток інструментів автоматизованого підтвердження або спростування автентичності мультимедійних матеріалів.

Додатковим фактором актуальності є стрімкий прогрес генеративних моделей, що забезпечують створення діпфейків із високою фотореалістичністю, узгодженістю освітлення та текстур, а також переконливою синхронізацією міміки й артикуляції. На практиці це означає, що традиційне «візуальне» виявлення підробок людиною або прості евристичні правила (пошук артефактів, неузгоджень у контурі обличчя, типових помилок) втрачають ефективність, а отже зростає потреба у формальних і відтворюваних методах аналізу, здатних працювати при різних умовах зйомки та якості відео.

Особливо гостро проблема проявляється у доменно нестабільних сценаріях: діпфейки відрізняються за технологіями генерації, якістю компресії, роздільною здатністю, типом облич (вік, стать, етнічні особливості), наявністю часткових перекриттів (окуляри, маски), рухом

камери й фоном. Це зумовлює падіння якості моделей, натренованих на обмежених наборах даних або на «лабораторних» прикладах, і вимагає розроблення підходів, які спираються на більш стійкі ознаки — зокрема на формалізовані характеристики обличчя, його геометрію та просторово-часову динаміку, що менш чутливі до стилістичних і компресійних спотворень.

В умовах гібридних інформаційних загроз та систематичного застосування дезінформації зростає значущість методів, здатних забезпечити оперативне виявлення синтетичного контенту як на рівні окремих користувачів, так і на рівні організаційних процесів (модерація, моніторинг, OSINT-аналіз, цифрова експертиза). Для України ця потреба має особливий вимір, оскільки відеоматеріали використовуються як інструмент психологічного тиску, дискредитації інституцій і формування хибних наративів; тому надійність та швидкість алгоритмів детекції діпфейків прямо пов’язана з інформаційною стійкістю та захистом суспільної довіри.

Не менш важливою є й прикладна складова проблеми: ефективний метод розпізнавання діпфейків має бути інтегровним у програмні рішення, працювати за обмежених ресурсів, забезпечувати відтворюваність результатів і мати зрозумілі критерії якості (accuracy, precision/recall, F1-score), оскільки саме ці характеристики визначають можливість використання системи в реальних сценаріях — від попереднього скринінгу контенту до підтримки прийняття рішень у кібербезпеці та судово-експертній практиці. У цьому контексті дослідження, орієнтоване на поєднання нейромережевого підходу з формалізованими ознаками обличчя та евристичними критеріями аномальності, є практично значущим шляхом підвищення надійності й стабільності моделі.

Нарешті, актуальність визначається і науковим запитом на побудову більш інтерпретованих та стійких моделей детекції синтетичного відео, які не лише демонструють високу точність «на тесті», а й зберігають працездатність при зміні умов, пояснюють природу помилки та дають можливість підсилення системи за рахунок багатоканальних ознак (структура landmarks, мікроміка,

узгодженість рухів). Саме тому дослідження, яке фокусується на формалізації поведінкових та геометричних характеристик обличчя у відеопотоці та їх використанні в нейромережевій класифікації діпфейків, є своєчасним і необхідним для розвитку прикладних рішень інформаційної безпеки та цифрової довіри.

Метою кваліфікаційної роботи є розроблення нейромережевого методу розпізнавання діпфейків на основі формалізованих ознак обличчя та створення програмного модуля ідентифікації таких ознак у відеоданих із використанням сучасних методів комп’ютерного зору та глибинного навчання.

Для досягнення поставленої мети в роботі необхідно розв’язати такі основні завдання:

1) провести аналіз сучасних методів і технологій виявлення діпфейків, з особливим акцентом на підходах, орієнтованих на аналіз facial features та просторово-часової динаміки обличчя;

2) визначити найбільш інформативні ознаки обличчя для задачі ідентифікації фейкових відео та підходи до їхньої формалізації;

3) спроєктувати архітектуру програмного модуля, який забезпечує детекцію обличчя у відеокадрах, виділення ключових точок (facial landmarks) та формування ознакового простору для подальшої класифікації;

4) реалізувати програмний модуль, що здійснює обробку відео, екстракцію характеристик обличчя та класифікацію відеофрагментів за ознакою достовірності/підробки;

5) провести експериментальні дослідження на відкритих наборах даних реальних і синтетичних відео (зокрема Deepfake Detection Challenge) з оцінюванням точності, повноти, F1-міри та інших показників якості класифікації;

6) виконати порівняльний аналіз нейромережевого підходу із евристичним критерієм, що ґрунтується на аномаліях у структурі та поведінці обличчя, і оцінити можливості їхньої комбінованої інтеграції в єдину систему розпізнавання діпфейків.

Висновки до розділу 1

1. Діпфейки формують багатовимірну загрозу інформаційній безпеці, що охоплює політичну, економічну, соціальну та правову сфери: вони підривають довіру до цифрових медіа, полегшують кіберзлочинність і репутаційні атаки, ускладнюють використання відео- та аудіоматеріалів як надійних доказів і сприяють появі феномену «постправди», коли автентичний контент сприймається як потенційний фейк.

2. Огляд сучасних підходів до виявлення діпфейків показує, що найвищих результатів досягають глибинні моделі, зокрема згорткові мережі, гібридні CNN–LSTM–Transformer та відео-трансформери, які поєднують аналіз текстур, просторово-часової динаміки, міміки, Facial Action Units та контексту сцени; водночас гострою проблемою залишається чутливість до компресії, дисбалансу даних та зміни типів генеративних моделей.

3. Узагальнення проаналізованих робіт підтверджує, що аналіз ознак обличчя (геометрія, симетрія, наявність ключових компонентів, міміка) є стійкою та перспективною основою для детекції діпфейків: порушення природних пропорцій і поведінкових патернів важче «замаскувати» навіть складним генеративним моделям, що обґрунтовує вибір у даній роботі нейромережевого методу розпізнавання діпфейків, орієнтованого саме на формалізовані характеристики обличчя.

2 ПРОЕКТУВАННЯ МОДУЛЯ ІДЕНТИФІКАЦІЇ ОЗНАК ОБЛИЧ У ДІПФЕЙКАХ

2.1 Архітектура методу розпізнавання діпфейків на основі аналізу обличчя

Запропонований метод розпізнавання діпфейків ґрунтується на багаторівневому аналізі відеоданих, що реалізує послідовну обробку кадрів, виявлення обличчя, локалізацію його ключових ознак та евристичне оцінювання ймовірності підробки.

Архітектура методу описує концептуальну структуру обробки інформації та взаємодію функціональних етапів між собою. Основною метою методу є автоматичний аналіз і класифікація відеороликів за рівнем достовірності на основі характеристик обличчя.

Архітектура методу розпізнавання діпфейків на основі аналізу обличчя має трирівневу структуру (рисунок 2.1), до складу якої входять:

1. Рівень збору та попередньої підготовки даних. На цьому рівні здійснюється завантаження відеофайлів, зчитування метаданих, перетворення даних у придатний для нейромережевої обробки формат, а також формування репрезентативних кадрів і кліпів. Відео нормалізуються за форматом, кадровою частотою та роздільністю, виконується первинна фільтрація за технічними параметрами і перевірка відповідності відеофайлів метаданим. На цьому етапі також виконується базова детекція обличчя й вирізання областей з обличчям, які надалі подаються на вхід нейромережевих блоків.

2. Рівень детекції та ідентифікації ознак обличчя. Основною функцією цього рівня є виявлення об’єктів у кадрі із застосуванням попередньо навчених нейромережевих моделей детекції. Для кожного кадру здійснюється пошук типових facial-компонентів — фронтального обличчя, очей та профільного вигляду. На основі просторових координат знайдених елементів формується множина аналітичних ознак, що відображають геометричну та симетричну структуру обличчя.

3. Рівень аналізу та візуалізації результатів. На цьому етапі виконується евристичне оцінювання достовірності зображення на основі виявлених ознак та їх поєднань. Якщо в кадрі спостерігаються численні аномалії (наприклад, відсутність очей за наявності обличчя, наявне лише профільне обличчя без фронтального вигляду, декілька облич одночасно там, де очікується одне тощо), відео позначається як потенційно фейкове. Результати роботи методу відображаються графічно у вигляді обмежувальних рамок навколо знайдених областей та текстових підписів на кадрах, що вказують тип виявлених об’єктів та виявлені відхилення. Такий спосіб подання дає змогу користувачу візуально зіставити вихідне зображення з отриманими позначками та інтерпретувати висновки методу щодо достовірності відеофрагмента.

Рисунок 2.1 – Трирівнева архітектура методу розпізнавання діпфейків

Функціональна реалізація методу відображається через сукупність модулів, що відповідають окремим етапам обробки. Модуль обробки даних (Data Analysis Unit) виконує завантаження відеофайлів, їх валідацію та попередню конверсію до формату, придатного для подальшого аналізу. Модуль детекції ознак (Object Detection Unit) відповідає за пошук ключових областей інтересу з використанням класифікаторів, які зокрема локалізують обличчя, очі та профільні компоненти. Модуль аналізу підробки (Deepfake Evaluation Unit) формує набір евристичних показників, які кількісно характеризують рівень потенційної фальсифікації відео, спираючись на виявлені аномалії у структурі та конфігурації обличчя. Модуль візуалізації (Visualization Unit) забезпечує наочне відображення результатів роботи методу у вигляді відміток на кадрах та текстових повідомлень, що дає змогу інтерпретувати отримані висновки та використовувати їх у подальшому аналізі або прийнятті рішень.

Таким чином, архітектура запропонованого методу забезпечує логічну послідовність від отримання відео до автоматизованого визначення фальсифікації. Методологічно він базується на класичних алгоритмах комп’ютерного зору з евристичною надбудовою для оцінки підозрілих ознак, що робить його ефективним для базового виявлення діпфейків при обмежених обчислювальних ресурсах.

2.2 Математичні методи виділення та формалізації ознак обличчя

На основі описаної архітектури (підрозділ 2.1) сформовано математичну модель методу виділення та формалізації ознак обличчя. Вона визначає принципи отримання, параметризації та аналітичного опису характеристик,

які використовуються для розпізнавання діпфейкових відео. Ідентифікація ознак обличчя є одним із ключових етапів нейромережевого методу розпізнавання діпфейків, оскільки саме структура обличчя, геометричні співвідношення між його елементами та поведінкові закономірності містять ознаки потенційної фальсифікації. На даному етапі формується набір аналітичних параметрів, які надалі використовуються для оцінювання достовірності відеофрагмента. У межах нейромережевого підходу детекція обличчя та отримання структурних ознак здійснюється за допомогою попередньо навчених моделей, здатних локалізувати ключові компоненти та визначати їхні просторові координати з високою точністю.

Основою процесу є процедура отримання ключових точок обличчя (facial landmarks), які описують контури очей, носа, рота, підборіддя та інших структурних елементів. Множина ключових точок для одного обличчя має вигляд:

$$L = \{(x_i, y_i), | i = 1,2, \dots, K\}, \tag{2.1}$$

де K — кількість визначених точок;

(x_i, y_i) — координати i-тої точки у просторі кадру.

Отримані координати використовуються для побудови геометричних ознак, які відображають пропорції та симетричність обличчя. Однією з таких ознак є вектор попарних відстаней між ключовими точками:

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}, \tag{2.2}$$

що дає змогу охарактеризувати структуру обличчя через відносні зміни та деформації. Сукупність усіх таких відстаней формує вектор геометричних характеристик:

$$D = \{ d_{ij} | 1 \le i \le j \le K \}, \quad (2.3)$$

Окремо оцінюється горизонтальна симетрія обличчя, яка є чутливою до артефактів діпфейку. Для цього визначається вісь симетрії $x = x_c$, де x_c — середнє значення абсцис точок обличчя:

$$x_c = \frac{1}{K} \sum_{i-1}^{K} x_i \, , \quad (2.4)$$

Далі обчислюється відхилення симетрії:

$$s = \frac{1}{K} \sum_{i-1}^{K} |x_i - (2x_c - x_i)|, \quad (2.5)$$

яке дорівнює нулю для ідеально симетричного обличчя та зростає за наявності аномалій у штучно модифікованих відеофрагментах.

Крім геометричних параметрів, визначаються структурні ознаки, пов’язані з наявністю або відсутністю ключових facial-компонентів. Для цього перевіряється множина:

$$C = \{C_{eyes}, C_{face}, C_{profile}\}, \quad (2.6)$$

де кожен елемент множини є бінарним індикатором

$$c_j = \begin{cases} 1, \text{компонент виявлено} \\ 0, \text{ компонент відсутній} \end{cases}, \quad 2.7)$$

Відсутність або деформація facial-компонентів (наприклад, очей) є типовими аномаліями для багатьох моделей генерації діпфейків.

Формування підсумкового вектора ознак для одного кадру виконується шляхом об’єднання кількох груп параметрів:

$$F = [D,\ S,\ C], \tag{2.8}$$

де D — геометричні ознаки;

S — показники симетрії;

C — наявність структурних компонентів.

Отриманий вектор ознак F = [D, S, C] є узагальненим описом обличчя, який може бути використаний як вхідний вектор для подальшого навчання класифікаційної моделі.

У межах запропонованого підходу класифікаційна модель реалізується у вигляді логістичної регресії, яку доцільно інтерпретувати як найпростіший одношаровий нейромережевий класифікатор. Формально вихід моделі визначається виразом:

$$\hat{y} = \sigma(w^{\top}F + b), \tag{2.9}$$

де F – вектор ознак обличчя;

w – вектор вагових коефіцієнтів;

b – зміщення (bias);

σ(·) – сигмоїдна функція активації.

Таким чином, задача розпізнавання діпфейка зводиться до навчання параметрів w, b за множиною розмічених прикладів (REAL / FAKE), щоб значення $\hat{y}$ наближалося до ймовірності належності відеофрагмента до класу фейкових відео. Така інтерпретація дозволяє розглядати запропонований метод як нейромережевий, що працює на формалізованих ознаках обличчя.

Виявлення асиметрій, пропусків facial-компонентів або нестандартних пропорцій свідчить про можливу штучну трансформацію обличчя, що робить ці ознаки високочутливими до маніпуляцій діпфейкового характеру.

Запропоновані математичні методи виділення та формалізації ознак обличчя забезпечують побудову інваріантного простору параметрів, який є основою для нейромережевого класифікатора. Використання таких формалізованих геометричних та структурних характеристик підвищує стійкість методу до варіацій пози, освітлення і дозволяє ефективно детектувати синтетичні артефакти, властиві діпфейковим відео.

2.3 Алгоритм реалізації методу розпізнавання діпфейків на основі ознак обличчя

2.3.1 Підготовка вхідних відеоданих

На початковому етапі здійснюється завантаження відеофайлу та валідація його структури (формат, наявність кадрів, тривалість). Далі з відео виділяється один або декілька репрезентативних кадрів, які використовуються для аналізу. Кожен кадр нормалізується за роздільністю та перетворюється у відтінки сірого, що зменшує обчислювальне навантаження і спрощує подальшу обробку без втрати просторової структури обличчя (відповідні кроки відображено у верхній частині рисунка 2.2).

2.3.2 Виділення ознак обличчя, очей і профілю

На цьому етапі виконується пошук трьох типів об’єктів у кадрі:

- обличчя (face) – основна область інтересу, у межах якої здійснюється подальший аналіз;
- очі (eyes) – пара локальних компонентів, що є важливими для оцінювання симетрії та цілісності обличчя;

- профіль (profile) – боковий ракурс обличчя, наявність якого може свідчити про нетипову конфігурацію сцени.

Для кожного з цих компонентів визначаються прямокутні області (bounding boxes) з відповідними координатами. На основі положення та взаєморозташування областей "обличчя–очі–профіль" формуються геометричні ознаки (відстані, пропорції), показники симетрії та структурні індикатори наявності/відсутності кожного компонента. Сукупність цих характеристик узагальнюється у векторі ознак $F = [D,\ S,\ C]$ структура якого схематично подана на рисунку 2.3.

Рисунок 2.2 – Схема алгоритму реалізації методу розпізнавання діпфейків

Рисунок 2.3 - Структурна схема виділення ознак обличчя, очей і профілю

2.3.3 Евристичний аналіз аномалій

Сформований вектор ознак F використовується для первинного евристичного аналізу на предмет наявності аномалій, характерних для діпфейкових відео. Зокрема, перевіряються такі ситуації:

- виявлено обличчя без очей (компонент face присутній, а eyes відсутні);

- у кадрі є лише профільне обличчя без фронтального, хоча за контекстом очікується фронтальний ракурс;
- одночасно виявлено декілька облич там, де очікується одна особа;
- порушені типові пропорції між обличчям та очима або спостерігається значна асиметрія.

Для кожної такої аномалії визначається евристична вага, а їх сукупність формує інтегральний показник підозрілості. Якщо цей показник перевищує заданий поріг, відео на евристичному рівні класифікується як потенційно фейкове (рисунок 2.2).

2.3.4 Класифікація на основі нейромережевого підходу

Після евристичного аналізу вектор ознак використовується для навчання класифікаційної моделі на основі логістичної регресії, яка інтерпретується як простий одношаровий нейромережевий класифікатор. Вхідним шаром у такій інтерпретації є компоненти вектора (геометрія обличчя та очей, показники симетрії, бінарні індикатори наявності обличчя, очей і профілю), а вихідним нейроном – скалярна оцінка ймовірності належності відео до класу FAKE. Класифікаційне рішення приймається пороговим способом: якщо значення ймовірності перевищує заданий поріг, відео позначається як діпфейк, інакше – як справжнє. Цей етап реалізує нейромережеву інтерпретацію запропонованого методу, зберігаючи при цьому прозорість та інтерпретованість моделі (рисунок 2.2).

2.3.5 Подання результатів

На виході алгоритму формуються як числові, так і візуальні результати. До числових належать класифікаційна мітка (REAL / FAKE), значення показника підозрілості та метрики якості. Візуальні результати включають кадри з обмежувальними рамками навколо обличчя, очей та профільних компонентів, а також текстові позначки виявлених аномалій. Це забезпечує

можливість як автоматизованої обробки, так і подальшого експертного аналізу.

Таким чином, розроблений алгоритм реалізації методу, що поєднує поетапну обробку відеоданих, формалізоване виділення ознак обличчя, очей і профілю та нейромережеву класифікацію на основі логістичної регресії, забезпечує практичну основу для надійного розпізнавання діпфейкових відео.

Висновки до розділу 2

1. Спроєктовано цілісну архітектуру методу розпізнавання діпфейків, яка реалізує логічну послідовність переходу від завантаження відеоданих до прийняття класифікаційного рішення. Архітектура включає рівні попередньої обробки, детекції facial-компонентів та підсумкової інтерпретації результатів, що забезпечує узгоджену роботу всіх модулів системи.

2. Сформовано математичну модель виділення ознак обличчя, що ґрунтується на формалізації геометричних параметрів, показників симетрії та структурних індикаторів наявності ключових компонентів (обличчя, очей, профілю). Побудовано узагальнений вектор ознак $F=[D,S,C]F=[D,S,C]F=[D,S,C]$, який забезпечує інваріантний опис обличчя й дозволяє виявляти аномалії, характерні для діпфейкових відео. Показано, що логістичну регресію доцільно інтерпретувати як простий одношаровий нейромережевий класифікатор.

3. Розроблено алгоритм реалізації методу, що поєднує поетапну обробку відео, автоматичне виділення facial-ознак і подальшу класифікацію на основі нейромережевої моделі. Алгоритм забезпечує як числове рішення (REAL/FAKE), так і візуальне представлення виявлених аномалій, що підвищує інтерпретованість і практичну придатність запропонованої системи для виявлення діпфейкових відео.

3 РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНЕ ОЦІНЮВАННЯ НЕЙРОМЕРЕЖЕВОГО МЕТОДУ РОЗПІЗНАВАННЯ ДІПФЕЙКІВ

3.1 Програмна реалізація методу розпізнавання діпфейків на основі ознак обличчя

Програмна реалізація запропонованого методу розпізнавання діпфейків виконана мовою Python із використанням бібліотек OpenCV, pandas, scikit-learn та допоміжних інструментів для обробки відео та аналізу даних.

На першому етапі реалізовано підсистему завантаження та попередньої обробки відео. Для зчитування відеопотоків використовується клас cv.VideoCapture, який забезпечує покадровий доступ до вмісту файлу. Після успішного відкриття відео виконується валідація його структури (наявність кадрів, коректний формат, ненульова тривалість). Для подальшого аналізу з відео виділяється один або кілька репрезентативних кадрів (frozen frames), які конвертуються у формат градацій сірого, що знижує обчислювальне навантаження та спрощує подальшу обробку без втрати просторової структури (рисунок А.1).

Другим кроком є реалізація компонента виявлення обличчя та пов’язаних з ним об’єктів. Для цього створено окремий клас ObjectDetector, який інкапсулює завантаження попередньо навчених каскадів Гаара та виклик функцій детекції (рисунок А.2). Методи цього класу забезпечують одночасний пошук кількох типів об’єктів: фронтальних облич, профільних ракурсів та очей. Узагальнена функція detect_objects приймає на вхід кадр у відтінках сірого, застосовує відповідні класифікатори та повертає координати знайдених областей інтересу (bounding boxes), які надалі використовуються як вхідні дані для формування ознак (рисунок А.3). На візуальному рівні виявлені області позначаються прямокутниками та колами, що дає змогу оперативно оцінити коректність спрацювання детектора.

Окрему увагу в реалізації приділено інтеграції з метаданими набору Deepfake Detection Challenge (DFDC). Метадані містять інформацію про походження відео (original), їх належність до класів REAL/FAKE та структуру набору. У даній роботі ці дані використовуються для формування підвибірок: наприклад, групи фейкових відео, згенерованих на основі одного оригіналу, що дозволяє досліджувати сталі та варіативні ознаки підробки (рисунок А.4).

Наступним кроком реалізовано функцію detect_fake_signs, яка автоматизує первинну евристичну перевірку відеозаписів на предмет наявності характерних ознак можливих діпфейків (рисунок А.5). Логіка функції базується на аналізі кількості та конфігурації знайдених облич, очей та профільних ракурсів, а також на виявленні аномальних поєднань (наприклад, обличчя без очей, лише профіль при очікуванні фронтального вигляду тощо). Виявлені ситуації фіксуються у вигляді текстових описів і надалі трансформуються в формалізовані ознаки.

Ключовим елементом програмної реалізації є функція extract_features_from_video, яка формує структурований опис одного відеофрагмента на основі його першого кадру (рисунок А.6). Функція виконує детекцію облич, очей і профілю, обчислює кількість кожного типу об’єктів, оцінює можливу асиметрію за положенням очей і формує інтегральний показник підозрілості (suspicion_score). Паралельно будується список словесних характеристик (issues), які пояснюють, які саме аномалії були виявлені. Результат роботи функції представлений у вигляді словника з числовими та категоріальними полями, придатними для конвертації у табличний формат.

Для масштабування аналізу на множину відеофайлів розроблено функцію build_suspicion_dataset, яка організовує пакетну обробку колекції відео (рисунок А.7). На вхід подається список шляхів до файлів, для кожного з яких послідовно викликається extract_features_from_video. Отримані результати об’єднуються в єдиний датафрейм з числовими ознаками (кількість облич, очей, профілів, показник підозрілості тощо) та текстовими полями, що описують виявлені проблеми. Для відстеження прогресу обробки

використовується модуль tqdm, що є важливим при роботі з великими відеоколекціями (рисунок А.8).

На завершальному етапі програмної реалізації підготовлений ознаковий датафрейм поєднується з «золотим стандартом» — справжніми мітками класів (REAL / FAKE), отриманими з метаданих змагання. Мітки кодуються у бінарному форматі (0 — справжнє відео, 1 — діпфейк), що дозволяє використовувати набір для побудови й оцінювання моделей класифікації.

Для евристичного методу, який приймає рішення на основі показника підозрілості (suspicion_score) та порогового правила, сформовано матрицю плутанини, що відображає співвідношення правильних і помилкових класифікацій (рисунок А.9). Додатково виконано побудову та оцінка моделі логістичної регресії на сформованих числових ознаках, а також отримано й проаналізовано звіт класифікації та прогнозні ймовірності належності до класу FAKE (рисунок А.10).

Як результат, програмна реалізація методу забезпечує повний цикл обробки: від читання відео та виділення структурних ознак обличчя до формування придатного для машинного навчання табличного набору, що є основою для подальшого кількісного оцінювання якості запропонованого методу розпізнавання діпфейків.

3.2 Дослідження та статистичний аналіз набору даних Deepfake Detection Challenge

Для експериментальної перевірки запропонованого методу використано фрагмент тренувального набору даних змагання Deepfake Detection Challenge (DFDC), метадані якого було попередньо проаналізовано та структуровано. На рисунку А.11 наведено узагальнену статистику цього підмножини, яка демонструє суттєвий дисбаланс класів: із 400 відеозаписів 323 (80,75%) мають мітку FAKE, тоді як частка справжніх відео (REAL) становить лише 19,25%. Усі записи належать до тренувальної частини набору (split = train), що підтверджує відсутність тестових зразків у наданому мета-файлі та відповідає

типовій організації змагальних датасетів, де тестова вибірка прихована організаторами (рисунок А.12).

Деталізований аналіз розподілу міток класів подано на рисунку А.13. Візуалізація підтверджує значну неврівноваженість даних на користь підроблених відео, що є характерним для задач детекції діпфейків, де кількість згенерованих прикладів, як правило, перевищує число автентичних. Така структура набору даних зумовлює необхідність уважного підходу до побудови та оцінювання класифікаційної моделі, зокрема використання стратифікованого розбиття на тренувальну й тестову вибірки, за потреби — введення ваг класів або застосування технік боротьби з дисбалансом (oversampling, undersampling, зміна функції втрат тощо).

Окремий аналіз було проведено для ідентифікації оригінальних відеоджерел, на основі яких генерувалися діпфейки. Виявлено, що найбільш поширене джерело — файл "meawmsgiti.mp4" — зустрічається лише 6 разів (1,858%), що свідчить про високу варіативність вихідних роликів та про розподілене походження фейкових зразків. Така різноманітність ускладнює задачу, оскільки модель має адаптуватися до широкого спектра сценаріїв зйомки, умов освітлення, ракурсів та індивідуальних особливостей облич.

Важливою складовою дослідження став візуальний аналіз перших кадрів відео. На рисунку А.14 продемонстровано вибірку кадрів із відеозаписів, класифікованих як FAKE. Вони дозволяють наочно оцінити типові артефакти, притаманні синтетичним відео: локальні викривлення текстур обличчя, розмиття контурів, аномалії в області очей та рота, неузгодженість освітлення між обличчям і фоном. На противагу цьому, рисунок А.15 ілюструє приклади кадрів зі справжніх відео (REAL), де спостерігаються природні варіації міміки, більш узгоджені тіні та відсутність явних ознак штучного «накладання» обличчя. Порівняння цих двох груп дає можливість виділити візуальні патерни, які надалі можуть бути формалізовані у вигляді ознак для автоматизованого аналізу.

Окремо було розглянуто випадок, коли декілька фейкових відео генеруються на основі одного оригінального запису. На рисунку А.16 наведено приклади кадрів з підроблених відео, що походять від спільного

джерела "meawmsgiti.mp4". Це дозволяє проаналізувати спектр трансформацій, які застосовуються до одного й того ж обличчя: зміну міміки, модифікацію руху губ, варіації в синхронізації з мовленням, різні ступені деформацій у периферійних зонах обличчя. Такий аналіз є важливим для розуміння того, як генеративні моделі комбінують сталі риси оригінального обличчя з синтетично зміненими компонентами, і які саме ознаки можуть бути найбільш інформативними для виявлення фальсифікації.

Проведене дослідження набору даних дозволяє зробити кілька висновків, важливих для подальшої побудови моделі. По-перше, суттєвий дисбаланс на користь класу FAKE вимагає ретельного вибору методів навчання та оцінювання, оскільки наївні стратегії можуть давати завищену оцінку якості за рахунок домінування одного класу. По-друге, висока різноманітність оригінальних відеоджерел і сценаріїв зйомки ускладнює задачу узагальнення, але одночасно робить результати моделі більш практично значущими у разі успішного навчання. По-третє, візуальний аналіз кадрів підтверджує наявність як грубих, так і тонких артефактів, пов’язаних з обличчям, очима та профілем, що узгоджується з вибраною в роботі стратегією фокусування на аналізі геометричних та структурних ознак обличчя як бази для нейромережевого методу розпізнавання діпфейків.

3.3 Експериментальна оцінка якості класифікації діпфейків

Експериментальна оцінка запропонованого методу передбачала порівняння двох підходів до класифікації відео: простого евристичного правила та моделі логістичної регресії, інтерпретованої як найпростіший нейромережевий класифікатор. В обох випадках класифікаційні рішення ухвалювалися на основі однакового ознакового простору, сформованого з кількості виявлених облич, очей, профільних об’єктів, показників асиметрії та інтегрального показника підозрілості (suspicion score). Це забезпечує коректність порівняння й дозволяє оцінити виграш саме від навчання моделі, а не від зміни ознак.

Перший підхід базується на фіксованому евристичному правилі: відео класифікується як фейкове, якщо suspicion score ≥ 2. Такий підхід є обчислювально дешевим і придатним для попереднього фільтрування великих колекцій відео в режимі наближеного скринінгу. За результатами експерименту для цього методу побудовано матрицю плутанини (рисунок 3.1), а також обчислено основні метрики якості: точність — 54,5%, прецизійність — 81,3%, повнота — 56,7%, F1-міра — 66,8%. Висока прецизійність за відносно низької загальної точності свідчить про те, що евристика добре «довіряється» своїм позитивним спрацюванням, але пропускає значну частину фейкових прикладів. Додаткову ілюстрацію дають окремі приклади відеофрагментів з обчисленими ознаками та виявленими аномаліями (рисунок 3.2).

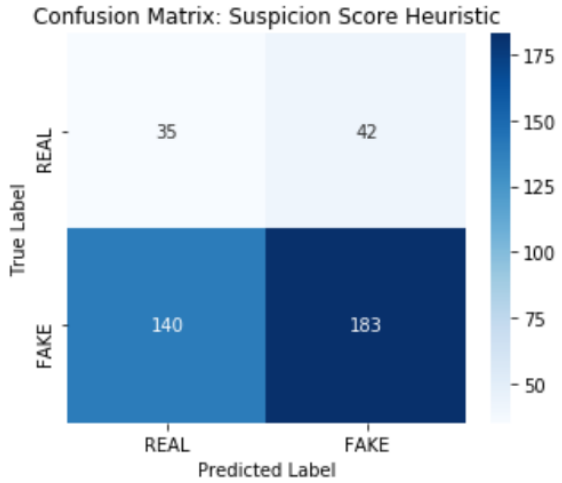


Рисунок 3.1 - Матриця плутанини та метрики для евристичного підходу

100%|██████████| 10/10 [00:02<00:00, 3.68it/s]

	video	n_faces	n_eyes	n_profiles	eye_asymmetry	suspicion_score	issues
0	eivxffliio.mp4	1	1	1	-1.000000	2	few_eyes
1	dwediigjit.mp4	1	4	1	-1.000000	0	
2	asvcrfdpnq.mp4	1	2	0	87.051709	1	eye_asymmetry
3	dntkzzzcdh.mp4	2	1	0	-1.000000	3	multiple_faces;few_eyes
4	dboxtiehng.mp4	1	3	1	-1.000000	0	
5	elginszwtk.mp4	0	0	0	-1.000000	4	no_face;few_eyes
6	cwbacdwrzo.mp4	1	2	0	95.036835	1	eye_asymmetry
7	afoovlsmtx.mp4	0	3	0	-1.000000	2	no_face
8	cmxcfkrjiv.mp4	0	1	0	-1.000000	4	no_face;few_eyes
9	aczrgyricp.mp4	1	4	0	-1.000000	0	

Рисунок 3.2 - Результати евристичного аналізу для 10 відеофрагментів

Другий підхід реалізовано у вигляді моделі логістичної регресії, навченої на сформованому наборі ознак. Усі відсутні значення в ознаковому просторі були замінені нульовими, після чого дані розділено на тренувальну та тестову вибірки у співвідношенні 80/20. Модель було натреновано на тренувальній підмножині й протестовано на відкладеній вибірці. Отримані результати засвідчили суттєве покращення якості класифікації порівняно з евристичним підходом: точність — 68,5%, прецизійність — 81%, повнота — 81%, F1-міра — 81% (рисунок 3.3).

🔎 Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	77
1	0.81	1.00	0.89	323
accuracy			0.81	400
macro avg	0.40	0.50	0.45	400
weighted avg	0.65	0.81	0.72	400

⚙️ Heuristic (Suspicion Score) Classification Report:

	precision	recall	f1-score	support
0	0.20	0.45	0.28	77
1	0.81	0.57	0.67	323
accuracy			0.55	400
macro avg	0.51	0.51	0.47	400
weighted avg	0.70	0.55	0.59	400

Рисунок 3.3 - Порівняння метрик логістичної регресії та евристики

Порівняльний аналіз показує, що при збереженні високої прецизійності модель значно підвищує повноту, тобто краще виявляє фейкові відео без критичного зростання кількості хибнопозитивних рішень. Приклади індивідуальних передбачень із зазначенням реальних міток та ймовірностей класу FAKE наведено на рисунку 3.4.

Важливим доповненням до кількісних метрик стали візуальні приклади роботи алгоритму на окремих кадрах відео (рисунок 3.5). Вони демонструють, як система локалізує обличчя, очі та профільні компоненти, а також фіксує аномалії у вигляді текстових діагностичних повідомлень. Зокрема, у випадках відсутності очей, наявності лише профільного обличчя або появи декількох облич у сцені відображаються відповідні індикатори, що корелюють із підвищеним значенням suspicion score та із рішенням класифікатора. Така

візуальна інтерпретованість є важливою перевагою методу, оскільки дозволяє експертам перевірити обґрунтованість рішень моделі.

	video	label_binary	predicted_fake	predicted_clf	predicted_proba
0	aagfhgtpmv.mp4	1	0	1	0.804130
1	aaprwogymq.mp4	1	1	1	0.832093
2	abarnvbtwb.mp4	0	1	1	0.725256
3	abofeumbvv.mp4	1	1	1	0.827438
4	abqwwspghj.mp4	1	0	1	0.785748
5	acifjvzvpm.mp4	1	1	1	0.822967
6	acqfdwsrhi.mp4	1	1	1	0.797575
7	acxnxvbsxk.mp4	1	0	1	0.719036
8	acxwigylke.mp4	1	0	1	0.719036
9	aczrgyricp.mp4	1	0	1	0.688206

Рисунок 3.4 - Порівняння класифікацій для прикладів відео

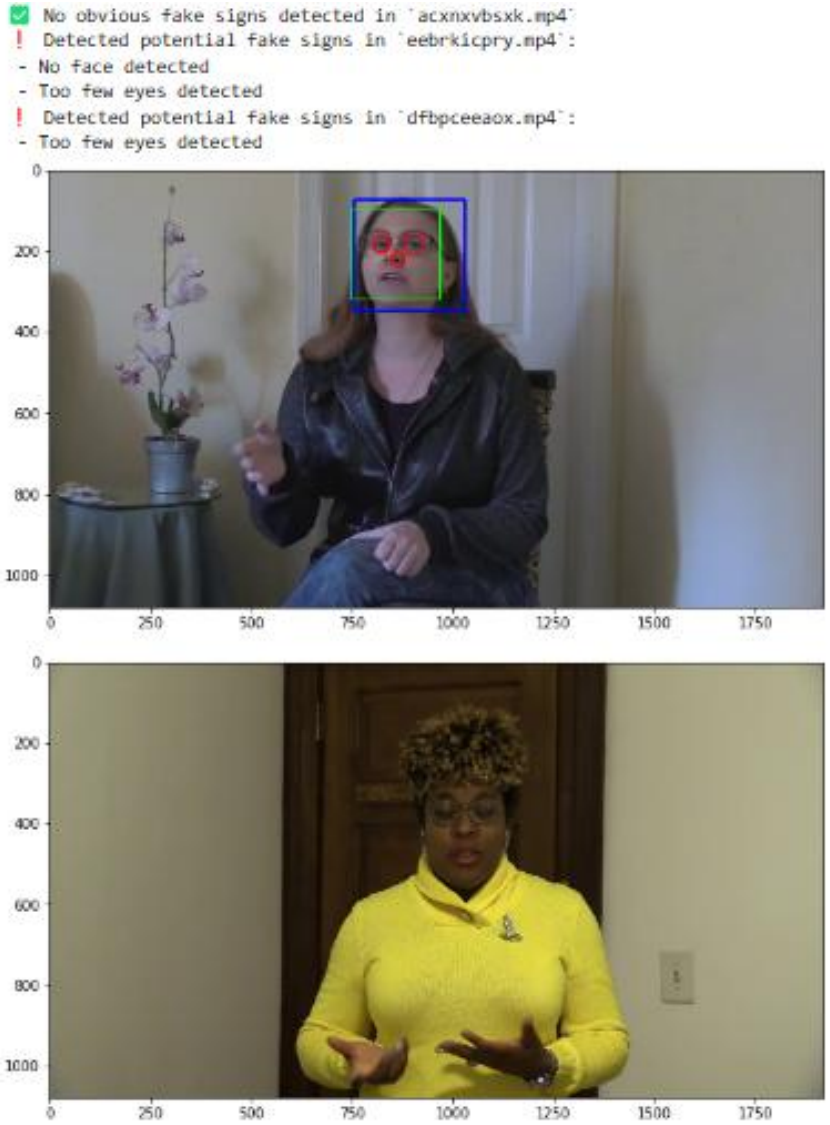


Рисунок 3.5 - Приклади фреймів відео з позначенням виявлених ознак та текстовими діагнозами

Узагальнюючи результати експериментів, можна зробити висновок, що запропонований підхід до класифікації діпфейків є доцільним у гібридній конфігурації: евристичний компонент доцільно використовувати як швидкий попередній фільтр, тоді як навчена на логістичній регресії модель виконує роль основного механізму прийняття рішень з вищою точністю та збалансованістю метрик. Така комбінація дозволяє одночасно забезпечити прийнятну продуктивність і достатній рівень надійності при виявленні діпфейкових відео.

3.4 Моделювання та порівняльний аналіз евристичного й нейромережевого підходів класифікації діпфейків

У цьому підрозділі наведено результати моделювання двох підходів до класифікації відеофрагментів: евристичного методу, що ґрунтується на пороговій оцінці показника підозрілості, та нейромережевого методу, реалізованого у вигляді логістичної регресії, інтерпретованої як одношаровий нейронний класифікатор. Обидва підходи працюють у спільному ознаковому просторі, сформованому на основі кількісних характеристик конфігурації обличчя, очей, профільних ракурсів, а також геометричних і симетричних параметрів.

Евристичний підхід передбачає обчислення інтегрального показника підозрілості (suspicion score), який формується на основі виявлених аномалій: відсутності очей за наявності обличчя, домінування профільного ракурсу без фронтального, наявності кількох облич у кадрі замість одного, значної асиметрії положення очей тощо. Класифікаційне рішення приймається за простим правилом: якщо suspicion score перевищує заздалегідь встановлений поріг, відео вважається фейковим, інакше — справжнім. Побудована матриця плутанини та відповідні метрики (точність, прецизійність, повнота, F1-міра) показують, що такий підхід забезпечує прийнятний рівень прецизійності, але характеризується обмеженою загальною точністю через значну кількість хибнонегативних класифікацій, що зумовлено відсутністю адаптації до даних.

Нейромережевий підхід класифікації реалізовано на основі логістичної регресії, яка розглядається як одношаровий перцептрон із сигмоїдною функцією активації вихідного нейрона. Вхідним вектором моделі виступає формалізований вектор ознак F=[D,S,C], що включає геометричні параметри, показники симетрії та бінарні індикатори наявності обличчя, очей і профільних компонентів. Перед навчанням пропущені значення замінюються нульовими, а вибірка даних розділяється на тренувальну та тестову частини у фіксованій пропорції. Після навчання модель генерує ймовірнісну оцінку

належності відео до класу FAKE, яка перетворюється на бінарне рішення за пороговим правилом. Порівняння метрик для двох підходів демонструє, що логістична регресія забезпечує вищу збалансованість між прецизійністю та повнотою, а також покращену загальну точність класифікації.

Отримані результати дають підстави стверджувати, що евристичний метод доцільно використовувати як швидкий механізм попереднього скринінгу, тоді як нейромережевий підхід на основі логістичної регресії слід розглядати як основний інструмент прийняття класифікаційного рішення. Такий гібридний сценарій дозволяє поєднати обчислювальну простоту та інтерпретованість евристики з вищою адаптивністю та якістю класифікації навченої моделі, що є важливим для практичних систем виявлення діпфейкових відео.

Висновки до розділу 3

1. Реалізовано повний програмний конвеєр розпізнавання діпфейків, який охоплює завантаження відео, покадрову обробку, детекцію обличчя, очей і профільних ракурсів, формування ознак та побудову структурованого ознакового датафрейму для подальшого машинного навчання.
2. Проведений статистичний аналіз набору DFDC підтвердив істотний дисбаланс класів та високу варіативність фейкових відео, що потребує обережного вибору методів навчання та обґрунтовує корисність створеного ознакового простору, побудованого на структурних характеристиках обличчя.
3. Експериментальні результати показали перевагу нейромережевого підходу (логістичної регресії) над евристичним правилом, оскільки модель забезпечила істотно вищу точність, збалансованість метрик та здатність виявляти більшу кількість діпфейкових відео без різкого зростання хибних спрацювань.

4. Гібридне поєднання евристики та навченої моделі виявилося оптимальним, оскільки дозволяє використовувати евристичний модуль як швидкий первинний фільтр, а нейромережевий класифікатор — як основний механізм ухвалення рішень, що підвищує надійність і практичну ефективність системи в реальних сценаріях виявлення діпфейків.

ВИСНОВКИ

Відповідно до поставлених у вступі завдань у роботі отримано такі результати:

1. Проведено ґрунтовний аналіз сучасних методів виявлення діпфейків, зокрема глибоких CNN (XceptionNet, LRNet), легковагових моделей MesoNet, гібридних архітектур CNN–LSTM–Transformer, аномалійно-орієнтованих моделей та відео-трансформерів ISTVT із урахуванням їхніх кількісних показників (AUC до 0,99 і вище) та стійкості до компресії й змін домену. Особливу увагу приділено підходам, орієнтованим на аналіз facial features, оптичного потоку, Facial Action Units та просторово-часової динаміки обличчя, що дозволило обґрунтувати вибір напрямку дослідження на основі формалізованих характеристик обличчя.

2. На основі аналізу архітектури методу та математичної постановки задачі визначено найбільш інформативні ознаки обличчя для виявлення діпфейків: кількість виявлених облич, очей та профільних об’єктів, індикатори асиметрії та інтегральний показник підозрілості (suspicion score). Для цих ознак побудовано формальний опис у вигляді аналітичних параметрів, що відображають геометричні та поведінкові закономірності на обличчі й дозволяють перевести візуальні спостереження у числовий ознаковий простір для подальшої класифікації.

3. Спроєктовано трирівневу архітектуру програмного модуля, що включає рівень збору та попередньої підготовки відеоданих, рівень детекції об’єктів (обличчя, очі, профільні компоненти, ключові точки) та рівень візуалізації й інтерпретації результатів. Запропонована структура забезпечує послідовний перехід від відеокадрів до формованого ознакового простору facial landmarks, підтримує модульність і розширюваність, а також дає змогу інтегрувати розроблений модуль у склад більш масштабних систем аналізу відеоконтенту.

4. Реалізовано повноцінний програмний модуль, який автоматизує увесь цикл обробки: читання та декомпозицію відео на кадри, виявлення й локалізацію обличчя та його ключових ознак, обчислення числових характеристик (кількість облич, очей, профілів, інтегральний suspicion score), формування ознакового датафрейму та злиття його із «золотим стандартом» міток REAL/FAKE. Отриманий табличний набір даних адаптовано для подальшого навчання та оцінювання класифікаційних моделей, а модуль візуалізації забезпечує наочне відображення результатів і текстових діагностичних повідомлень на кадрах відео.

5. Проведено експериментальні дослідження на фрагменті тренувального набору Deepfake Detection Challenge (400 відео, з яких 80,75 % мають мітку FAKE), із розрахунком ключових метрик – точності, прецизійності, повноти та F1-міри. Для евристичного критерію на основі порогу suspicion score ≥ 2 отримано accuracy 54,5 %, precision 81,3 %, recall 56,7 % і F1-міру 66,8 %, що демонструє високу «обережність» правила при суттєвій втраті частини фейкових прикладів; для нейромережевого класифікатора (логістичної регресії) точність зросла до 68,5 %, а precision, recall та F1-міра досягли 81 %, що підтверджує ефективність запропонованого ознакового простору й доцільність використання навчуваної моделі.

6. Виконано порівняльний аналіз евристичного підходу та нейромережевої моделі, які працюють на спільному ознаковому просторі facial features, що забезпечило коректність зіставлення отриманих результатів. Показано, що логістична регресія забезпечує значно кращий баланс між точністю й повнотою та виявляє більше діпфейкових відео, тоді як евристика зберігає високу прецизійність за рахунок пропуску частини фейків; на цій основі обґрунтовано гібридну схему інтеграції, де евристичний модуль виконує роль швидкого первинного фільтра, а нейромережевий класифікатор – основного механізму ухвалення рішень, що підвищує надійність і практичну ефективність системи розпізнавання діпфейків.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Retrieved from https://arxiv.org/abs/1901.08971

2. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7). IEEE. https://doi.org/10.1109/WIFS.2018.8630761

3. Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2021). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Retrieved from https://arxiv.org/abs/2008.12262

4. Janūtėnas, R., Blažauskas, T., & Damaševičius, R. (2023). Deep learning methods to detect image falsification. In R. Damaševičius & W. Wei (Eds.), *ICT Systems and Sustainability* (pp. 97–112). Springer. https://doi.org/10.1007/978-3-031-34601-1_7

5. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2021). Towards explainable deepfake detection: An optical flow based CNN method for unmasking deepfakes. *Pattern Recognition Letters*, *146*, 41–47. https://doi.org/10.1016/j.patrec.2021.02.016

6. Nassif, A. B., & Shahin, I. (2022). Improved optical flow estimation for deepfake detection. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13276-3

7. Tan, T., Zhang, S., Li, Y., Wang, S., & Zhou, J. (2023). Facial Action Dependencies Estimation for DeepFake Video Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/25348

8. Petmezas, G., Voulodimos, A., Doulamis, A., & Doulamis, N. (2024). Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20548-6

9. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., & Li, S.-N. (2024). Explore and enhance the generalization of anomaly deepfake detection. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2406.04940

10. Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., & Tang, J. (2023). ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 1335–1348. https://doi.org/10.1109/TIFS.2023.3239223

11. Coccomini, D. A., Messina, N., & Gennaro, C. (2022). *Combining efficientnet and vision transformers for video deepfake detection*. Springer. https://link.springer.com/chapter/10.1007/978-3-031-06433-3_19

12. Petmezas, G., Vanian, V., & Konstantoudakis, K. (2025). *Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification*. https://link.springer.com/article/10.1007/s11042-024-20548-6

13. Liu, P., Tao, Q., & Zhou, J. T. (2024). *Evolving from Single-modal to Multi-modal Facial Deepfake Detection: A Survey*. https://arxiv.org/abs/2406.06965

14. Salvi, D., Liu, H., Mandelli, S., Bestagini, P., & Zhou, W. (2023). *A robust approach to multimodal deepfake detection*. Journal of Imaging, 9(6), 122. https://doi.org/10.3390/jimaging9060122

15. Kaddar, B., Fezza, S. A., Akhtar, Z., & Hamidouche, W. (2024). *Deepfake detection using spatiotemporal transformer*. https://doi.org/10.1145/3643030

16. Wang, J., Wu, Z., Ouyang, W., Han, X., & Chen, J. (2022). *M2tr: Multi-modal multi-scale transformers for deepfake detection*. https://doi.org/10.1145/3512527.3531415

17. Soudy, A. H., Sayed, O., Tag-Elser, H., & Ragab, R. (2024). *Deepfake detection using convolutional vision transformers and convolutional neural networks*. https://link.springer.com/article/10.1007/s00521-024-10181-7

18. Wang, Z., Cheng, Z., Xiong, J., Xu, X., & Li, T. (2024). *A timely survey on vision transformer for deepfake detection*. https://arxiv.org/abs/2405.08463

19. Hashmi, A., Shahzad, S. A., & Lin, C. W. (2025). *AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection*. https://ieeexplore.ieee.org/abstract/document/10938399/

20. VP, S. E., & Dheepthi, R. (2024). *Llm-enhanced deepfake detection: Dense cnn and multi-modal fusion framework for precise multimedia authentication*. https://ieeexplore.ieee.org/abstract/document/10533511/

21. Jung, T., Kim, S., & Kim, K. (2020). *DeepVision: Deepfakes detection using human eye blinking pattern*. https://ieeexplore.ieee.org/abstract/document/9072088/

22. Liu, W., She, T., Liu, J., Li, B., & Yao, D. (2024). *Lips Are Lying: Spotting the Temporal Inconsistency between Audio and Visual in Lip-Syncing DeepFakes*. https://proceedings.neurips.cc/paper_files/paper/2024/hash/a5a5b0ff87c59172a13342d428b1e033-Abstract-Conference.html

23. Agarwal, S., & Farid, H. (2021). *Detecting deep-fake videos from aural and oral dynamics*. https://openaccess.thecvf.com/content/CVPR2021W/WMF/html/Agarwal_Detecting_Deep-Fake_Videos_From_Aural_and_Oral_Dynamics_CVPRW_2021_paper.html

24. Mazaheri, G. (2022). *Detection and localization of facial expression manipulations*. http://openaccess.thecvf.com/content/WACV2022/html/Mazaheri_Detection_and_Localization_of_Facial_Expression_Manipulations_WACV_2022_paper.html

25. Chakraborty, R., & Naskar, R. (2024). *Role of human physiology and facial biomechanics towards building robust deepfake detectors: A comprehensive*

survey and analysis. https://www.sciencedirect.com/science/article/pii/S1574013724000613

26. Waseem, S., Bakar, S. A. R. S. A., Ahmed, B. A., & Omar, Z. (2023). *DeepFake on face and expression swap: A review*. https://ieeexplore.ieee.org/abstract/document/10285057/

27. Zamboni, J. (2025). *Deepfake detection via facial landmark motion analysis during person-specific word pronunciation*. https://reposit.haw-hamburg.de/handle/20.500.12738/16769

28. Agarwal, S. (2021). *Detecting synthetic faces by understanding real faces*. https://farid.berkeley.edu/downloads/publications/sathesis21.pdf

29. Saif, S., Tehseen, S., & Ali, S. S. (2024). *Fake news or real? Detecting deepfake videos using geometric facial structure and graph neural network*. https://www.sciencedirect.com/science/article/pii/S0040162524002671

30. Hashmi, A., Shahzad, S. A., Lin, C. W., & Tsao, Y. (2024). *Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights*. https://arxiv.org/abs/2411.07650

31. Lipianina-Honcharenko, K., Lendiuk, D., Ivaniush, A., Boguta, G., Soia, M., Yurkiv, K. An Intelligent Method for Recognizing Real and Generated Ukrainian-Language Voices. In 2024 5th IEEE KhPI Week on Advanced Technology (KhPIWeek 2024), Kharkiv, Ukraine, October 2024. IEEE Conference Proceedings, DOI: 10.1109/KHPIWEEK61434.2024.10877988.

32. Lipianina-Honcharenko, K., Yarych, V., Ivasechko, A., Filinyuk, A., Yurkiv, K., Lebid, T. Evaluating the Effectiveness of Attention-Gated-CNN-BGRU Models for Historical Manuscript Recognition in Ukraine. In CEUR Workshop Proceedings, 1st International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development (AISD 2024), Ternopil, Ukraine, 10–11 May 2024.

33. Lipianina-Honcharenko, K., Soia, M., Yurkiv, K., Ivasechko, A. Evaluation of the Effectiveness of Machine Learning Methods for Detecting Disinformation in Ukrainian Text Data. In CEUR Workshop Proceedings, 7th

International Workshop on Computer Modeling and Intelligent Systems (CMIS 2024), Zaporizhzhia, Ukraine, 3 May 2024.

34. Lipianina-Honcharenko, K., Ihnatiev, I., Bohuta, H., Yurkiv, K., & Novosad, S. (2025). Rule-based method for aligning media content with ukrainian legislation.

35. Lipianina-Honcharenko, K., Komar, M., Bohuta, H., Ihnatiev, I., Yurkiv, K., Illiashenko, O., & Bilovus, L. (2025). A general method for real-time detection of information threats with a Ukraine case study. *Radioelectronic and Computer Systems*, *2025*(3), 202-230.

36. Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів. Тернопіль: ЗУНУ, 2024. 83 с.

37. Комар М.П., Саченко А.О., Васильків Н.М., Загородня Д.І. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за другим (магістерським) рівнем вищої освіти. – Тернопіль: ЗУНУ, 2024. – 32 с.

Додаток А

Візуалізація етапів виявлення діпфейків

```
def display_image_from_video(video_path):
    capture_image = cv.VideoCapture(video_path)
    ret, frame = capture_image.read()
    fig = plt.figure(figsize=(10,10))
    ax = fig.add_subplot(111)
    frame = cv.cvtColor(frame, cv.COLOR_BGR2RGB)
    ax.imshow(frame)
```

Рисунок А.1 - Завантаження відео та обробка кадрів

```
class ObjectDetector():

    def __init__(self,object_cascade_path):

        self.objectCascade=cv.CascadeClassifier(object_cascade_path)

    def detect(self, image, scale_factor=1.3,
               min_neighbors=5,
               min_size=(20,20)):

        rects=self.objectCascade.detectMultiScale(image,
                                                  scaleFactor=scale_factor,
                                                  minNeighbors=min_neighbors,
                                                  minSize=min_size)
        return rects
```

Рисунок А.2 - Використання каскадів Гаара для виявлення об'єктів

```
def detect_objects(image, scale_factor, min_neighbors, min_size):

    image_gray=cv.cvtColor(image, cv.COLOR_BGR2GRAY)

    eyes=ed.detect(image_gray,
                   scale_factor=scale_factor,
                   min_neighbors=min_neighbors,
                   min_size=(int(min_size[0]/2), int(min_size[1]/2)))

    for x, y, w, h in eyes:
        cv.circle(image,(int(x+w/2),int(y+h/2)),(int((w + h)/4)),(0, 0,255),3)

    # deactivated due to many false positive
    #smiles=sd.detect(image_gray,
    #                 scale_factor=scale_factor,
    #                 min_neighbors=min_neighbors,
    #                 min_size=(int(min_size[0]/2), int(min_size[1]/2)))

    #for x, y, w, h in smiles:
    #    #detected smiles shown in color image
    #    cv.rectangle(image,(x,y),(x+w, y+h),(0, 0,255),3)


    profiles=pd.detect(image_gray,
                   scale_factor=scale_factor,
                   min_neighbors=min_neighbors,
                   min_size=min_size)

    for x, y, w, h in profiles:
        cv.rectangle(image,(x,y),(x+w, y+h),(255, 0,0),3)

    faces=fd.detect(image_gray,
                   scale_factor=scale_factor,
                   min_neighbors=min_neighbors,
                   min_size=min_size)

    for x, y, w, h in faces:

        cv.rectangle(image,(x,y),(x+w, y+h),(0, 255,0),3)

    fig = plt.figure(figsize=(10,10))
    ax = fig.add_subplot(111)
    image = cv.cvtColor(image, cv.COLOR_BGR2RGB)
    ax.imshow(image)
```

Рисунок А.3 - Інтеграція детекторів і побудова візуалізації

```
same_original_fake_train_sample_video = list(meta_train_df.loc[meta_train_df.original=='kgbkktcjxf.mp4'].index)
for video_file in same_original_fake_train_sample_video[1:4]:
    print(video_file)
    extract_image_objects(video_file)
```

```
byqzyxifza.mp4
cwrtyzndpx.mp4
cwwandrkus.mp4
```

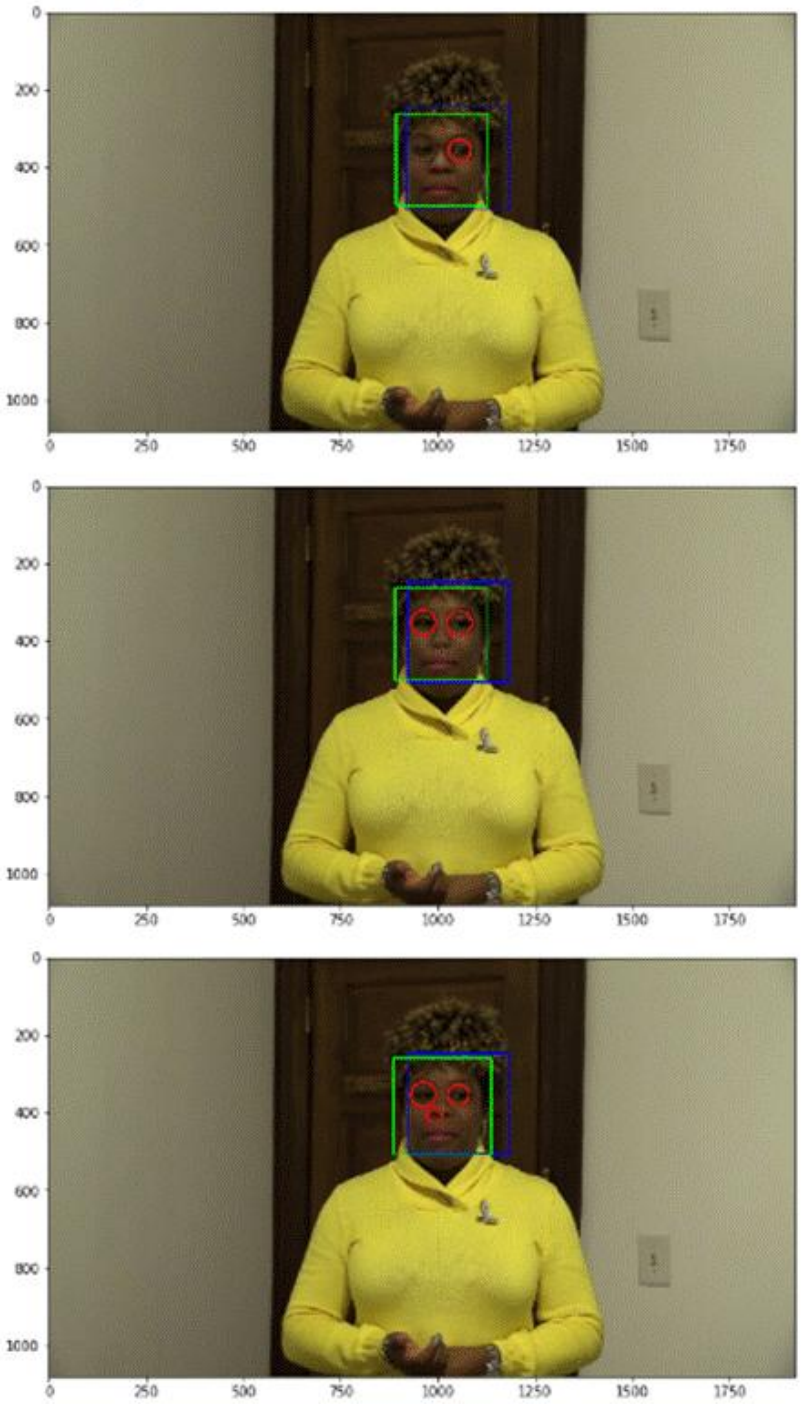


Рисунок А.4 - Обробка відео та інтеграція з метаданими створених на основі одного оригіналу

```python
def detect_fake_signs(video_file, video_set_folder=TRAIN_SAMPLE_FOLDER):

    video_path = os.path.join(DATA_FOLDER, video_set_folder, video_file)
    capture_image = cv.VideoCapture(video_path)
    ret, frame = capture_image.read()
    if not ret:
        print(f"Couldn't read frame from {video_file}")
        return

    gray = cv.cvtColor(frame, cv.COLOR_BGR2GRAY)

    faces = fd.detect(gray, scale_factor=1.3, min_neighbors=5, min_size=(50, 50))
    eyes = ed.detect(gray, scale_factor=1.3, min_neighbors=5, min_size=(20, 20))
    profiles = profile_detector.detect(gray, scale_factor=1.3, min_neighbors=5, min_size=(50, 50))

    issues = []

    if len(faces) == 0:
        issues.append("No face detected")
    elif len(faces) > 1:
        issues.append("Multiple faces detected")

    if len(eyes) < 2:
        issues.append("Too few eyes detected")
    elif len(eyes) > 4:
        issues.append("Too many eye-like regions detected")

    if len(profiles) > 0 and len(faces) == 0:
        issues.append("Profile face detected without frontal face")

    detect_objects(frame.copy(), scale_factor=1.3, min_neighbors=5, min_size=(50, 50))

    if issues:
        print(f"❗ Detected potential fake signs in `{video_file}`:")
        for issue in issues:
            print(" -", issue)
    else:
        print(f"✅ No obvious fake signs detected in `{video_file}`")

for video_file in fake_train_sample_video[:3]:
    detect_fake_signs(video_file)
```

Рисунок А.5 - Функція detect_fake_signs, яка виконує автоматизовану первинну перевірку відеозаписів на наявність ознак підробки

```python
def extract_features_from_video(video_file, video_set_folder=TRAIN_SAMPLE_FOLDER):

    video_path = os.path.join(DATA_FOLDER, video_set_folder, video_file)
    capture_image = cv.VideoCapture(video_path)
    ret, frame = capture_image.read()
    if not ret:
        return {'video': video_file, 'issue': 'Cannot read frame'}

    gray = cv.cvtColor(frame, cv.COLOR_BGR2GRAY)

    faces = fd.detect(gray, scale_factor=1.3, min_neighbors=5, min_size=(50, 50))
    eyes = ed.detect(gray, scale_factor=1.3, min_neighbors=5, min_size=(20, 20))
    profiles = profile_detector.detect(gray, scale_factor=1.3, min_neighbors=5, min_size=(50, 50))

    score = 0
    issues = []

    if len(faces) == 0:
        score += 2
        issues.append("no_face")
    elif len(faces) > 1:
        score += 1
        issues.append("multiple_faces")

    if len(eyes) < 2:
        score += 2
        issues.append("few_eyes")
    elif len(eyes) > 4:
        score += 1
        issues.append("too_many_eyes")

    if len(profiles) > 0 and len(faces) == 0:
        score += 1
        issues.append("profile_only")

    eye_asymmetry = -1
    if len(eyes) == 2:
        x1, y1, _, _ = eyes[0]
        x2, y2, _, _ = eyes[1]
        dx = abs(x1 - x2)
        dy = abs(y1 - y2)
        eye_asymmetry = np.sqrt(dx**2 + dy**2)
        if eye_asymmetry > 80:  # евристика
            score += 1
            issues.append("eye_asymmetry")

    return {
        'video': video_file,
        'n_faces': len(faces),
        'n_eyes': len(eyes),
        'n_profiles': len(profiles),
        'eye_asymmetry': eye_asymmetry,
        'suspicion_score': score,
        'issues': ";".join(issues)
    }
```

Рисунок А.6 - Функція виявлення ознак підробки у відео

```python
def build_suspicion_dataset(video_list, folder=TRAIN_SAMPLE_FOLDER):
    feature_rows = []
    for video_file in tqdm(video_list):
        features = extract_features_from_video(video_file, folder)
        if features is not None:
            feature_rows.append(features)
    return pd.DataFrame(feature_rows)
```

Рисунок А.7 – Функція створення ознакового набору з відео

```python
from tqdm import tqdm

video_list_10 = train_list[:10]

df_10 = build_suspicion_dataset(video_list_10, folder=TRAIN_SAMPLE_FOLDER)

import pandas as pd
from IPython.display import display

display(df_10)
```

```
100%|██████████| 10/10 [00:02<00:00,  3.68it/s]
```

	video	n_faces	n_eyes	n_profiles	eye_asymmetry	suspicion_score	issues
0	eivxffliio.mp4	1	1	1	-1.000000	2	few_eyes
1	dwediigjit.mp4	1	4	1	-1.000000	0	
2	asvcrfdpnq.mp4	1	2	0	87.051709	1	eye_asymmetry
3	dntkzzzcdh.mp4	2	1	0	-1.000000	3	multiple_faces;few_eyes
4	dboxtiehng.mp4	1	3	1	-1.000000	0	
5	elginszwtk.mp4	0	0	0	-1.000000	4	no_face;few_eyes
6	cwbacdwrzo.mp4	1	2	0	95.036835	1	eye_asymmetry
7	afoovlsmtx.mp4	0	3	0	-1.000000	2	no_face
8	cmxcfkrjiv.mp4	0	1	0	-1.000000	4	no_face;few_eyes
9	aczrgyricp.mp4	1	4	0	-1.000000	0	

Рисунок А.8 –Формування датафрейму ознак для 10 відеофайлів

```python
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

all_video_files = list(meta_train_df.index)
df_all = build_suspicion_dataset(all_video_files, folder=TRAIN_SAMPLE_FOLDER)

df_all['label'] = df_all['video'].map(meta_train_df['label'])

df_all['label_binary'] = df_all['label'].map({'REAL': 0, 'FAKE': 1})

df_all['predicted_fake'] = df_all['suspicion_score'].apply(lambda x: 1 if x >= 2 else 0)

cm = confusion_matrix(df_all['label_binary'], df_all['predicted_fake'])

plt.figure(figsize=(5, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['REAL', 'FAKE'], yticklabels=['REAL', 'FAKE'])
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.title("Confusion Matrix: Suspicion Score Heuristic")
plt.show()
```

```
100%|██████████| 400/400 [01:53<00:00,  3.52it/s]
```

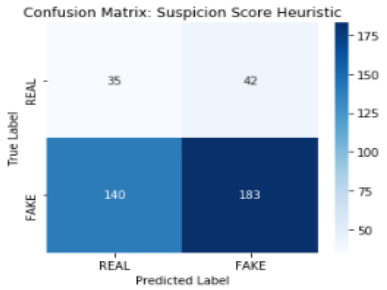


Рисунок А.9 –Матриця плутанини для евристичного методу на основі підозрілості

```python
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

feature_cols = ['suspicion_score', 'n_faces', 'n_eyes', 'n_profiles', 'eye_asymmetry']
X = df_all[feature_cols].fillna(0)
y = df_all['label_binary']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

clf = LogisticRegression(max_iter=1000)
clf.fit(X_train, y_train)

df_all['predicted_clf'] = clf.predict(X)
df_all['predicted_proba'] = clf.predict_proba(X)[:, 1]  # ймовірність FAKE

comparison_df = df_all[['video', 'label_binary', 'predicted_fake', 'predicted_clf', 'predicted_proba']]
display(comparison_df.head(10))

print("🔎 Logistic Regression Classification Report:")
print(classification_report(df_all['label_binary'], df_all['predicted_clf']))
print("\n⚙️ Heuristic (Suspicion Score) Classification Report:")
print(classification_report(df_all['label_binary'], df_all['predicted_fake']))
```

Рисунок А.10 –Побудова та оцінка моделі логістичної регресії для виявлення діпфейків

	label	split	original
Total	400	400	323
Most frequent item	FAKE	train	meawmsgiti.mp4
Frequence	323	400	6
Percent from total	80.75	100	1.858

Рисунок А.11 - Частотний розподіл значень ознак label, split та original у метаданих тренувального набору

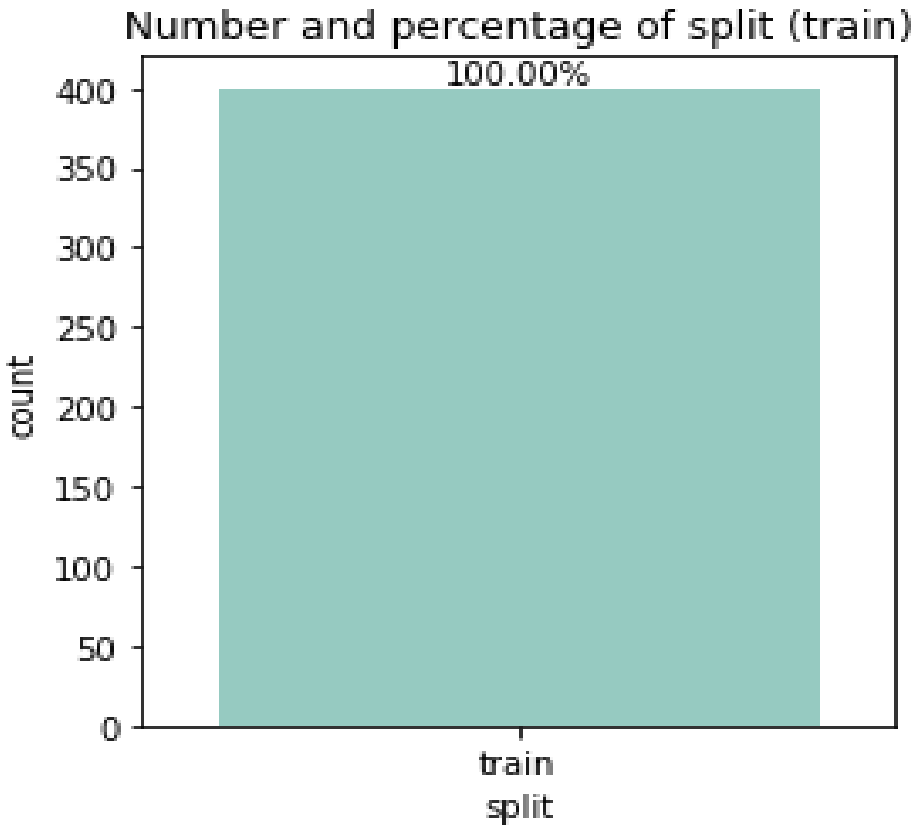


Рисунок А.12 - Первинний аналіз метаданих тренувального набору даних

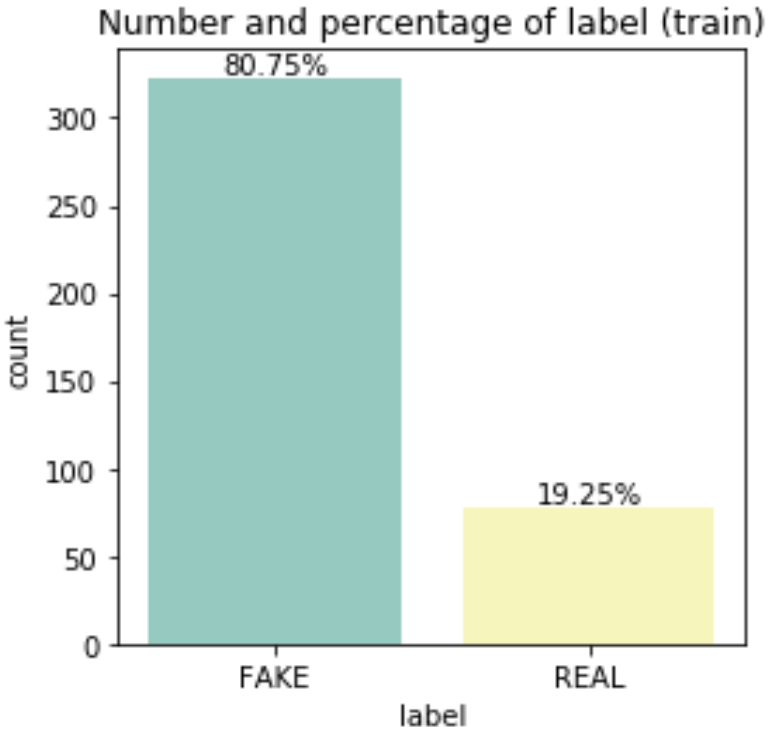


Рисунок А.13 - Розподіл міток в тренувальному наборі даних

Рисунок А.14 – Перші кадри з підроблених відео

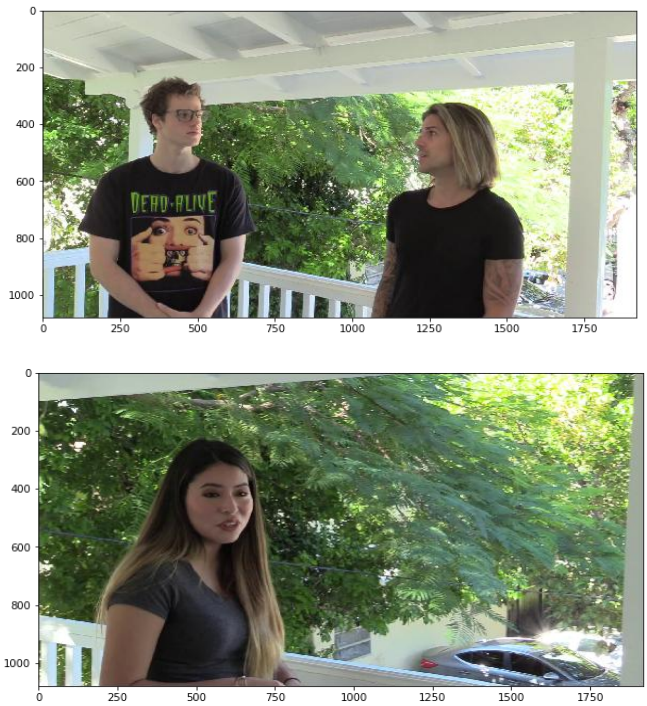


Рисунок А.15 - Кадри з справжніх відео

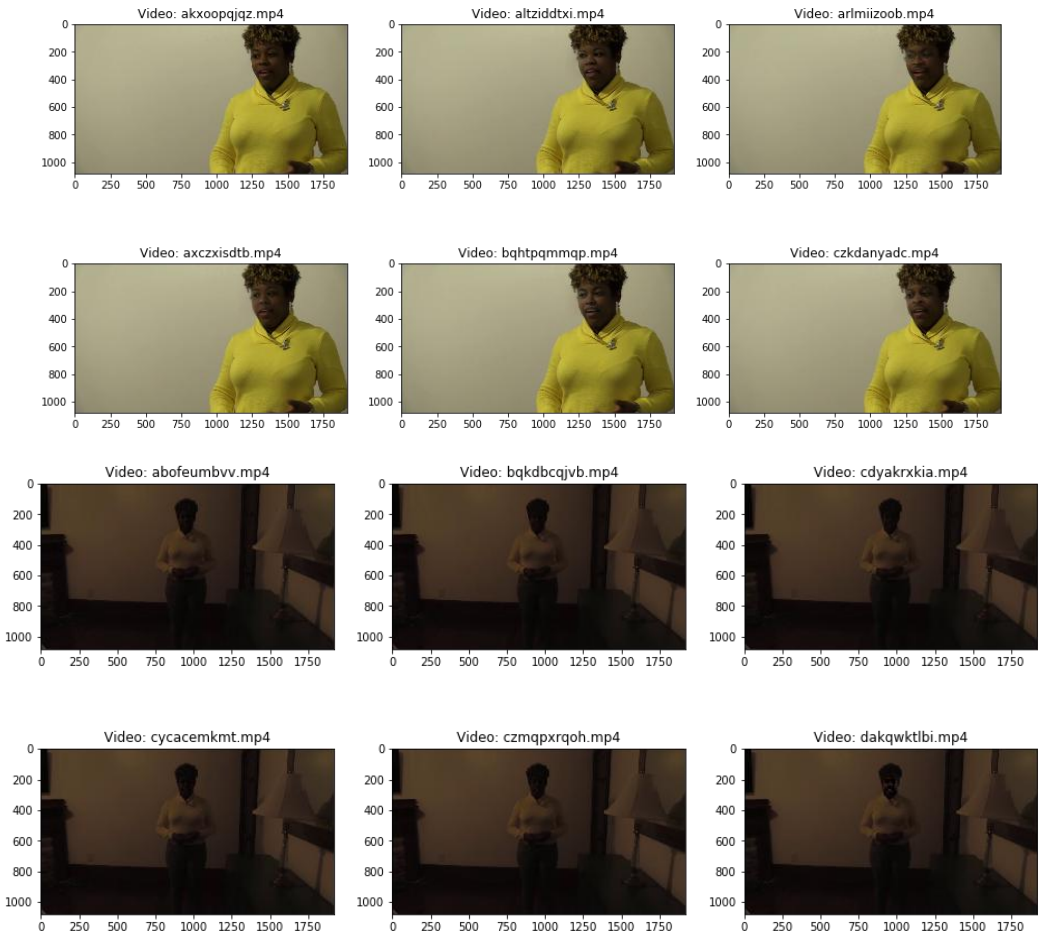
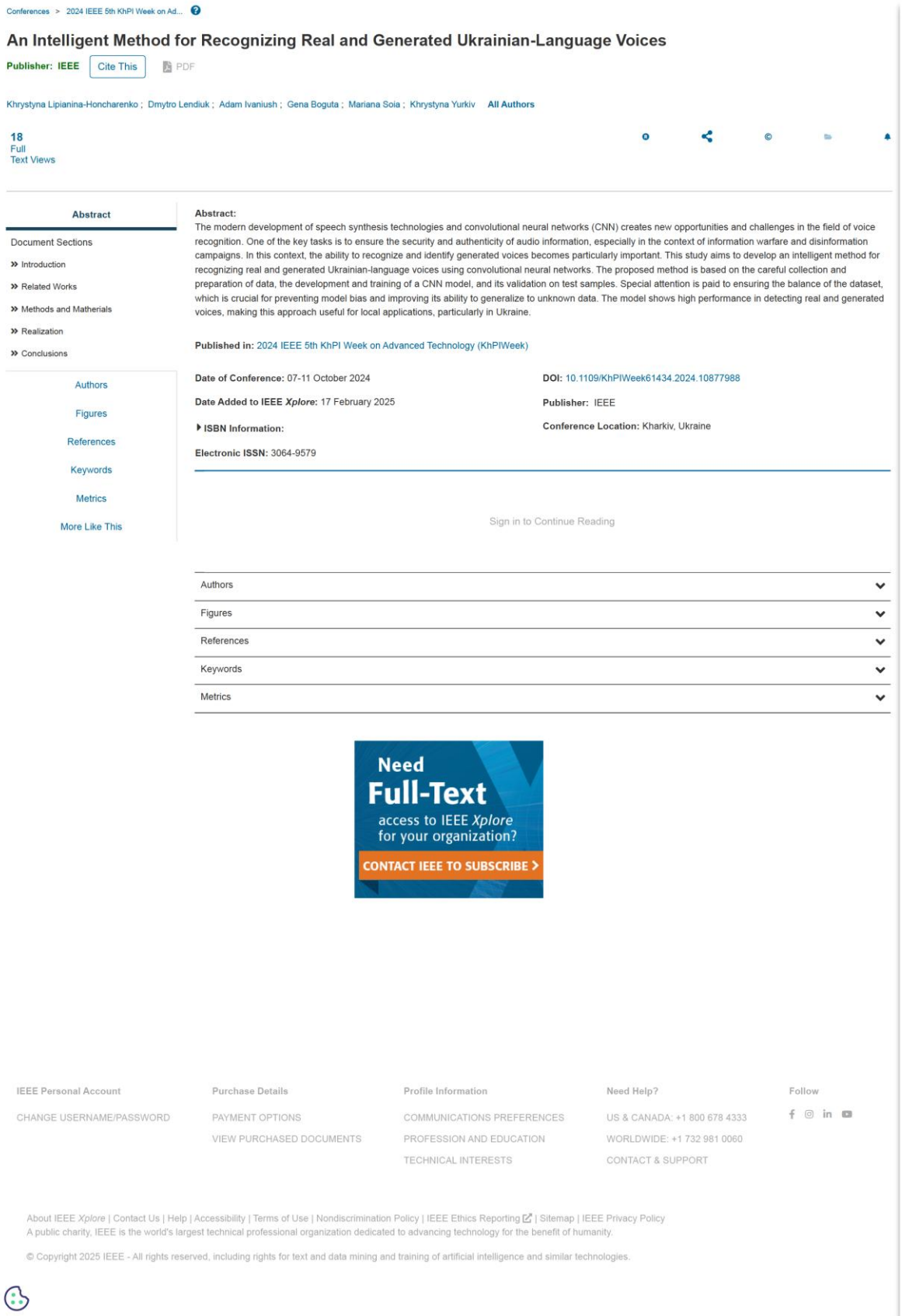


Рисунок А.16 - Кадри з підроблених відео, що походять з одних джерел

Додаток Б

Апробація отриманих результатів



Conferences > 2024 IEEE 5th KhPI Week on Ad...

An Intelligent Method for Recognizing Real and Generated Ukrainian-Language Voices

Publisher: IEEE | Cite This | PDF

Khrystyna Lipianina-Honcharenko ; Dmytro Lendiuk ; Adam Ivaniush ; Gena Boguta ; Mariana Soia ; Khrystyna Yurkiv **All Authors**

18
Full
Text Views

Abstract

Document Sections

» Introduction

» Related Works

» Methods and Matherials

» Realization

» Conclusions

Authors

Figures

References

Keywords

Metrics

More Like This

Abstract:
The modern development of speech synthesis technologies and convolutional neural networks (CNN) creates new opportunities and challenges in the field of voice recognition. One of the key tasks is to ensure the security and authenticity of audio information, especially in the context of information warfare and disinformation campaigns. In this context, the ability to recognize and identify generated voices becomes particularly important. This study aims to develop an intelligent method for recognizing real and generated Ukrainian-language voices using convolutional neural networks. The proposed method is based on the careful collection and preparation of data, the development and training of a CNN model, and its validation on test samples. Special attention is paid to ensuring the balance of the dataset, which is crucial for preventing model bias and improving its ability to generalize to unknown data. The model shows high performance in detecting real and generated voices, making this approach useful for local applications, particularly in Ukraine.

Published in: 2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek)

Date of Conference: 07-11 October 2024

DOI: 10.1109/KhPIWeek61434.2024.10877988

Date Added to IEEE *Xplore*: 17 February 2025

Publisher: IEEE

ISBN Information:

Conference Location: Kharkiv, Ukraine

Electronic ISSN: 3064-9579

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

Need Full-Text access to IEEE *Xplore* for your organization?

CONTACT IEEE TO SUBSCRIBE >

IEEE Personal Account
CHANGE USERNAME/PASSWORD

Purchase Details
PAYMENT OPTIONS
VIEW PURCHASED DOCUMENTS

Profile Information
COMMUNICATIONS PREFERENCES
PROFESSION AND EDUCATION
TECHNICAL INTERESTS

Need Help?
US & CANADA: +1 800 678 4333
WORLDWIDE: +1 732 981 0060
CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy
A public charity, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

Vol-3716
urn:nbn:de:0074-3716-4

AISD 2024
Artificial Intelligence for Sustainable Development 2024

Proceedings of the First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development

Ternopil, Ukraine, May 10-11, 2024.

Edited by

Mykola Dyvak*
Oleh Pitsun **

* West Ukrainian National University, Ternopil, Ukraine
** West Ukrainian National University, Ternopil, Ukraine

Table of Contents

2024-06-11: submitted by Oleh Pitsun, metadata incl. bibliographic data published under Creative Commons CC0
2024-07-02: published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073) |valid HTML5|

Evaluating the Effectiveness of Attention-Gated-CNN-BGRU Models for Historical Manuscript Recognition in Ukraine

Khrystyna Lipianina-Honcharenko[1], Volodymyr Yarych[1], Andrii Ivasechko[1], Anatoliy Filinyuk[1], Khrystyna Yurkiv[1], Tetian Lebid [1]

[1] *West Ukrainian National University, Lvivska str., 11, Ternopil, 46000, Ukraine*

Abstract

This research focuses on evaluating the effectiveness of the Attention-Gated-CNN-BGRU model for Handwritten Text Recognition from historical documents, specifically from the archive of Khmelnytskyi Oblast, written in Ukrainian and Russian languages between 1861 and 1913. The methodology involved preprocessing, data augmentation, deep learning with attention mechanism, and expert assessment. The obtained results showed an average percentage of correctly recognized characters at 71.7%, demonstrating the high effectiveness of the model. A strong negative correlation between text complexity and recognition accuracy underscores the need for further improvement in Optical Character Recognition technologies. The main direction of future research will be adapting the model for recognizing texts written in the Ukrainian language using the Latin alphabet, which is crucial for preserving Ukraine's cultural heritage.

Keywords

Historical documents, Optical Character Recognition, Handwritten Text Recognition, deep learning

1. Introduction

The modern world of digital technologies opens new horizons for the preservation and study of historical documents, offering unique tools for exploring cultural heritage. One of the key directions in this field is the development and application of OCR systems, which automate the process of converting image-based text into machine-readable format. This, in turn, facilitates easier access to historical documents, their analysis, and interpretation. However, the characteristics of historical manuscripts, such as variability in writing styles, degree of preservation, and material erosion, pose challenges for researchers that require the development of specialized OCR algorithms and methods.

This scientific article is dedicated to analyzing the effectiveness of the Attention-Gated-CNN-BGRU model [1], specifically developed for text recognition from manuscripts. The research is based on a carefully curated dataset of documents from the state archive of Khmelnytskyi Oblast, covering the years from 1861 to 1919 and various funds reflecting the socio-historical aspects of the region during the specified period. The main focus of the research is on determining the accuracy of the OCR model in the context of variability in manuscript complexity, evaluating the impact of writing styles and document preservation on recognition quality, as well as analyzing potential directions for algorithm optimization.

The First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development; May 10-11, 2024, Ternopil, Ukraine

); xrustya.com@gmail.com (K. Lipianina-Honcharenko); v.yarych@wunu.edu.ua (V. Yarych); andrewivasechko@gmail.com (A. Ivashechkoafilinyuk@ukr.net(A. Filinyuk) kh.yurkiv@wunu.edu.ua (K. Yurkiv); Liebid89@gmail.com (T. Lebid).

); 0000-0002-2441-6292 (K. Lipianina-Honcharenko); 0000-0003-4455-952X (V. Yarych); 0009-0002-8623-7828 (A. Ivashechko0000-0002-2659-114X (A. Filinyuk); 0009-0007-4917-3251 (K. Yurkiv); 0009-0005-3413-9064 (T. Lebid);

Given the importance of preserving historical heritage and the development of digital humanities, the results of this research aim not only to demonstrate the potential of modern OCR technologies in historical text analysis but also to provide valuable insights for further improvement of digital humanities tools.

The formulation of the problem. In the context of the development of digital technologies and the study of cultural heritage, there arises the problem of effectively recognizing text from historical manuscripts using OCR systems. This problem arises due to the diversity of writing styles, differences in the preservation state of documents and their materials, complicating the process of automated conversion of images into machine-readable form. Such technical challenges create the necessity for the development and enhancement of specialized OCR algorithms to effectively process historical manuscripts.

This article presents a method for evaluating the effectiveness of the Attention-Gated-CNN-BGRU model for text recognition from historical manuscripts, based on deep learning with attention mechanism. Chapter 2 discusses a review of related works; Chapter 3 describes the research methodology, including dataset description and document recognition approach; Chapter 4 is dedicated to the implementation of the algorithm and detailed analysis of the obtained results. Chapter 5 presents the research conclusions summarizing the main findings and emphasizing the prospects for further research in this field.

2. Related work

The development of artificial intelligence (AI) technologies is one of the key and significant trends of the present day. This is primarily explained by the ability of AI, or more accurately "computational" or "electronic" intelligence, to learn and self-learn, recognize and synthesize language and images. Worldwide practice already has results of AI activity in the field of historical research. For example, Swedish scientists were able to reconstruct a complex handwritten text from the 18th century, stored in the Swedish National Archives [2]. The object of this research was the decrees on freedom of the press from 1766, which were written in Latin script. Such breakthroughs in science prompt us to explore the capabilities of AI in interpreting handwritten texts from Ukrainian archives, which were written in Cyrillic.

In Ukraine, specialists from not only the Institute of Artificial Intelligence Problems of the Ministry of Education and Science of Ukraine and the National Academy of Sciences of Ukraine are working on the development and improvement of AI capabilities, but also scientific and pedagogical workers and researchers from departments and divisions of AI, computer science and applied mathematics, computer technologies and economic cybernetics, innovative technologies, intellectual information systems, mathematical modeling, AI systems, information security of many higher education institutions and research institutes of Ukraine. A leading role in this field is played by the team of the Research Institute of Intelligent Computer Systems of the Western Ukrainian National University (Ternopil) and the V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine (Kyiv).

Modern research in the field of AI is actively developing, especially in the area of document interpretation, where information technologies are used for analysis, search, and interpretation of relevant texts. Significant important informational potential for our research[3] useful information on the creation of archival collections of websites within the framework of initiative documentation at the Central State Electronic Archives is described in the research [4]. In research [5] characterized the process of digitizing archival documents in foreign countries. The topical issue of the role of digital sources in the research [6]. In research [7] predicted the growing role of archive digitization in modern society.

Innovative approaches to preserving cultural heritage through the use of image recognition and augmented reality, as seen in the researchs [8-10], illuminate the development of the intersection of technology and history. These methodologies resonate with the fundamental principles of our research, which applies Attention-Gated-CNN-BGRU models for recognizing historical manuscripts in Ukraine, demonstrating the versatility and potential of deep learning in

various fields. Similar to how [11, 12], utilized deep neural networks, and researchs [13, 14], applied multilevel data encoding, our study employs advanced machine learning algorithms to open new perspectives in the analysis and preservation of cultural heritage, affirming the critical role of technology in safeguarding and researching our collective past.

Research [15] explored offline handwritten text recognition (HTR) with reduced training datasets, proposing a model trained on crossed-out text to effectively recognize such words without compromising accuracy. In [16], methods for classifying handwritten words into a digital format were investigated, combining direct word classification using Convolutional Neural Networks (CNN) and character segmentation with Long Short-Term Memory (LSTM) networks. Additionally, the study in [17] addressed handwritten text conversion and storage in ASCII format, covering preprocessing, feature extraction, and classification using deep learning methods. The research in [18] enhanced HTR performance on lines with crossed-out words, while the approach in [19] proposed outperformed traditional OCR systems. The integration of HTR into OCR systems was discussed in [20], alongside outlining an HTR competition focusing on historical documents [21]. Effective practices in HTR, including basic architectures and datasets like IAM and RIMES, were described in [22], while the importance of image processing in handwritten text detection was emphasized in [23], showing potential in forgery protection and handwriting analysis. Lastly, the Attention-Gated-CNN-BGRU model [1] for Ukrainian HTR is freely available, trained on historical Ukrainian texts provided by researchers and libraries.

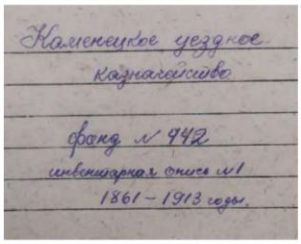
The Attention-Gated-CNN-BGRU model [1] for Ukrainian HTR is freely available, trained on historical Ukrainian texts and provided by researchers and libraries.The analysis of known solutions in the field of HTR, including the application of models like CRNN, and approaches to text classification and segmentation, underscores significant progress in this domain. However, this study focuses on analyzing the effectiveness of the Attention-Gated-CNN-BGRU model [1] for recognizing text from historical manuscripts, distinguishing itself through the use of attention mechanisms for detailed analysis of contextual relationships between characters. The approach in this research demonstrates improvements in recognizing complex historical texts, especially with crossed-out words and conditions of high text complexity, making it akin in concept to the works of Jose Carlos Aradillas et al. [15], but with an additional focus on adaptation to the specificity of Ukrainian handwritten text. This sets apart this study from others, providing a new approach to addressing the problem of text recognition in historical documents using deep learning and attention mechanisms, opening up prospects for further advancements in this field. The analysis of the text written in Ukrainian is given in the work [25-26].

3. Methodology

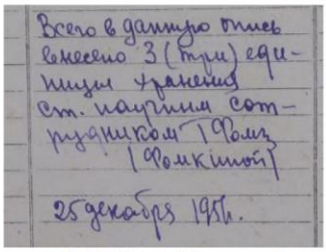
3.1. Dataset Description

For this study, documents of varying readability levels were collected from the State Archives of Khmelnytskyi Oblast. Among the proposed documents, digitized descriptions of ancient records from the following funds were taken: Fund 442, Kamianets-Podilskyi County Treasury for 1861-1913; Fund 507, Office of the Chief of the South-Western Customs District for 1907-1913; Fund 596, Podilia Branch of Princess Tetiana Mykolaivna's Committee for Providing Temporary Assistance to Victims of Military Actions for 1914-1915; Fund 598, Investigative Court of the 2nd Division of Kamianets-Podilskyi County for 1875-1880; Fund 616, Kamianets-Podilskyi County Military Affairs for 1884-1919; as well as Fund 309, Isakovets Customs for 1931-1915, written in Ukrainian and Russian languages.

The dataset [24] for recognizing ancient handwritten text consists of 75 PNG-format images, examples of which are shown in Figure 1.



A) Easy to read



B) Difficult to read

Figure 1: Examples of texts

3.2. Description of Document Recognition Approach

This research focuses on analyzing the effectiveness of the Attention-Gated-CNN-BGRU model for recognizing text from historical manuscripts. Modern challenges in OCR in historical documents include a variety of writing styles, document preservation levels, and character variability. This approach involves applying deep learning with attention to the context of characters and their interactions within words or text fragments, which enhances recognition accuracy compared to traditional methods. The implementation was done using the Python programming language. The approach consists of the following stages:

Stage 1. Data preprocessing:

1.1. Digital transformation: Each text sample I from historical manuscripts is converted into a digital format through scanning or photography, where I is represented as a matrix of pixels $I(x, y)$, where x and y are pixel coordinates.

1.2. Normalization: Applying normalization to each image I to standardize sizes and color intensities. The normalized image can be represented as

$$Inorm = f(I),$$

where f is the normalization function.

1.3. Data augmentation: Generating new samples Iaug from Inorm using transformations such as rotation, scaling, shifting, etc., to increase data diversity:

$$Iaug = g(Inorm),$$

where g is the augmentation function.

Stage 2. The Attention-Gated-CNN-BGRU model combines CNN for efficient visual feature extraction with gated recurrent units (GRU) and attention mechanism for modeling dependencies between characters. Step-by-step, this is presented as follows:

2.1. CNN: Using CNN to extract visual features $\phi(Iaug)$ from images, where ϕ is the feature extraction function implemented using CNN.

2.2. GRU and attention mechanism: Processing feature sequences $\phi(Iaug)$ using GRU and attention mechanism to model character dependencies and consider context:

$$\psi(\phi(Iaug)),$$

where ψ is the function implementing GRU and attention mechanism.

Stage 3. Model validation:

3.1. Word Error Rate (WER):

$$\text{WER}=\frac{S+D+I}{N}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the correct text.

3.2. Character Error Rate (CER):

$$CER = \frac{S + D + I}{M}$$

where M is the total number of characters in the correct text.

3.3. Levenshtein CER: Using the Levenshtein distance to calculate CER, where the Levenshtein distance between two strings is the minimum number of single-element edits (insertions, deletions, substitutions) required to transform one string into another.

Stage 4. Results analysis:

4.1. Performance evaluation: Analyzing the distribution of WER and CER errors to evaluate the model's effectiveness in recognizing text from historical manuscripts. Statistical methods are used to compare the model results with baseline indicators.
4.2. Impact of attention mechanism: Evaluating the contribution of the attention mechanism to the model's ability to identify complex characters and their contextual relationships through detailed analysis of corrected errors and improvements in recognition accuracy.
Further, this approach is described as an algorithm (Figure 2) for evaluating the effectiveness of the Attention-Gated-CNN-BGRU model for recognizing text from historical manuscripts, starting with data preparation, which includes data collection, digital transformation, and dataset augmentation from manuscripts. Next, model configuration involves integrating convolutional neural networks and gated recurrent units with attention mechanism for text analysis. Model validation is done using independent test datasets, and performance evaluation is based on word and character error rates. Results analysis includes comparison with baseline indicators, detailed recognition analysis, and identification of directions for further model improvement.

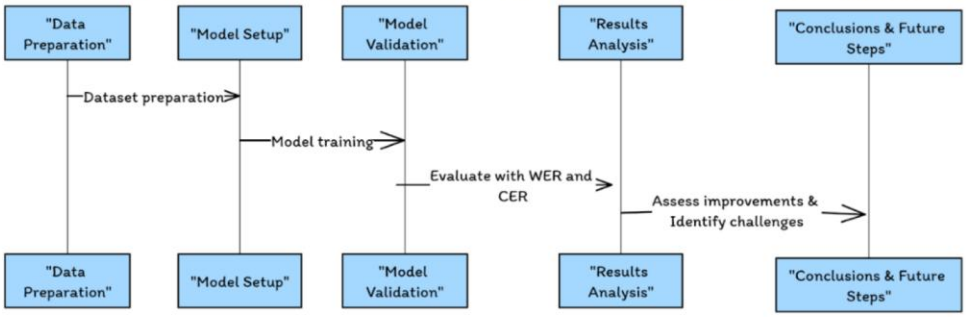


Figure 2. Structure of Document Recognition Approach

3.3 Approach to Expert Evaluation

Within the framework of this study, an expert evaluation methodology was used to assess the effectiveness of OCR models on historical manuscript materials. Five qualified experts analyzed the document samples, evaluating the difficulty of text recognition, the number of correctly recognized characters, the total number of recognized characters, and the total number of characters in each document. The approach proposed by the authors for this research for evaluation consists of the following stages:
Stage 1. Data preparation for analysis.
1.1. Collection of evaluations from experts ($EX1, EX2, EX3, EX4, EX5$) based on the following criteria:
1.1.1. Assessment of text recognition difficulty from 1 to 5, where 1 - very easy, 5 - very difficult.
1.1.2. Correctly recognized characters: the number of characters that were correctly recognized.
1.1.3. Recognized characters: the total number of recognized characters.
1.1.4. Total characters: the total number of characters in the original document.
1.2. Calculation of indicators:
1.2.1. Average assessment of text recognition difficulty ($SORT$):

$$SORT = \frac{\sum_{i=1}^{n} Assessment_i}{n}$$

where n is the number of experts, and Ratingi is the rating from expert i.
1.2.2. Average percentage of correctly recognized symbols ($SVPRS$):

$$SVPRS = \frac{\sum_{i=1}^{n} \left(\frac{Right\_recongised\_symbols_i}{All\_symbols} \times 100 \right)}{n}$$

1.2.3. Average number of recognized symbols (ANRS)

$$SKRS = \frac{\sum_{i=1}^{n} Recognised\_symbols_i}{n}$$

Stage 2. Analysis of the results:

2.1. Analysis of the overall effectiveness of the OCR model:

2.1.1. Calculating the average values for $SORT$, $SVPRS$, and $SKRS$ for all documents allows for the evaluation of the overall effectiveness of the OCR model.

2.1.2. Measuring the dispersion and standard deviation for each indicator helps understand the diversity of expert ratings and the variability of recognition results.

2.2. Impact of text complexity on recognition quality: Using correlation analysis between SORT and SVPRS helps determine how text complexity affects recognition quality. A positive correlation may indicate that as the complexity of the text increases, recognition efficiency decreases.

Stage 3. Interpretation of the obtained results:

3.1. The overall efficiency of the OCR model is determined through the analysis of the average values of $SORT$, $SVPRS$, and $SKRS$. High values of $SORT$ and $SKRS$ with low $AATRD$ indicate the model's ability to adapt to various conditions of handwritten text.

3.2. The results of analyzing the impact of text complexity on recognition quality provide insights into the limitations of the OCR model and directions for further improvement. Enhancing recognition accuracy under conditions of high text complexity may become a key aspect of model optimization.

The algorithm (Figure 3) for the experimental evaluation of OCR models on historical manuscripts begins with collecting expert ratings, where experts analyze text samples from manuscripts based on defined criteria such as text recognition complexity and the number of correctly recognized symbols. The second stage involves analyzing the overall efficiency of the models by calculating average values, dispersion, standard deviation of indicators, and conducting correlation analysis between text complexity and recognition quality. In the final stage, the obtained results are interpreted, allowing for the assessment of the overall efficiency of the models and determining optimization directions to improve recognition accuracy under conditions of high text complexity.

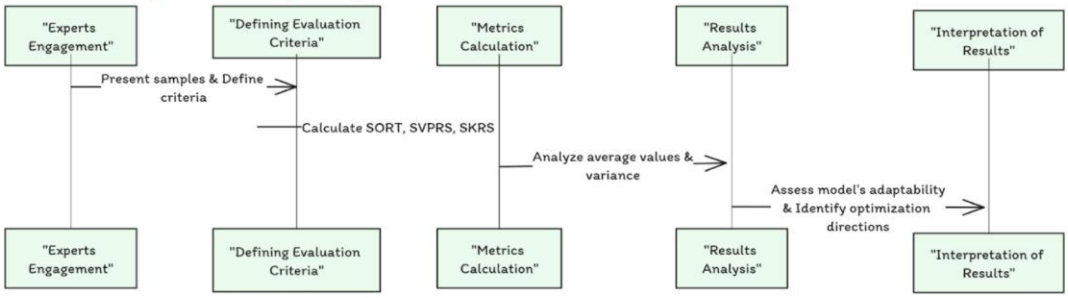


Figure 3. Structure of the OCR model effectiveness assessment approach

4. Result

In this section, a detailed analysis of the experiment results is presented, allowing for the evaluation of the effectiveness of applying OCR models to historical manuscripts, as described in section 3.1.

Table 1 below displays the results of recognizing the first five fragments of historical manuscripts using the OCR model. The data include links to the photos of the originals as well as the text recognized by the model. A comprehensive analysis of the entire dataset and detailed research results are available for review on GitHub [24], where an in-depth investigation of the OCR model's effectiveness on various fragments of historical documents is presented.

Upon completing the initial analysis of the text recognition results, a comprehensive expert evaluation was conducted, as detailed in section 3.3 of the documentation. The expertise involved

a thorough examination of text complexity, recognition quality, and symbol identification accuracy, aimed at gaining a deeper understanding of the effectiveness of applying OCR models to historical documents. Prominent experts involved in the analysis included Senior Lecturer of the Department of Information and Socio-Cultural Activities at the Western Ukrainian National University, Volodymyr Yarych, Doctor of Historical Sciences Anatoliy Filinyuk from Ivan Ogienko Kamianets-Podilskyi National University, as well as candidates of Historical Sciences Serhiy Sydoruk and Serhiy Trubchaninov. Additionally, Valentina Filinyuk, a candidate of Philological Sciences and Associate Professor at Khmelnytsky Humanitarian-Pedagogical Academy, contributed to the expert assessment. Their professional approach and deep knowledge facilitated the identification of key aspects for further improvement of OCR technologies in the context of working with historical texts.

Table 1.
Results of recognizing the first 5 fragments

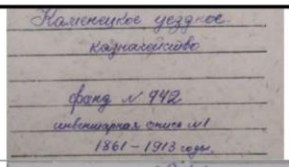
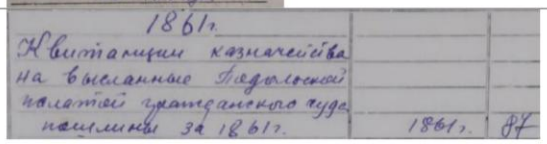
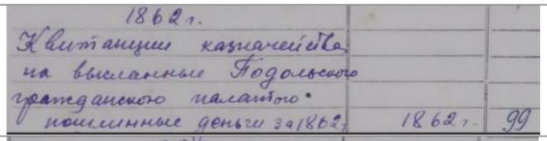
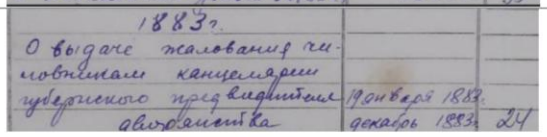
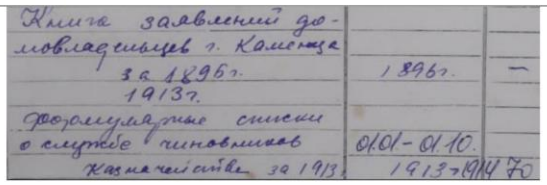
№ doc.	Link to the original photo:	Recognized text:
1		Каменецкое уездное. казначейство Фанд N 442 инвентарная опись л! 1861 – 1913 года.
2		Каннъ 1861г. Нвитанции казначенива на высланные Подолоской палатой гражданского чуда пошлины за 1861г. 1861г. 87
3		1862г. Хвитанции казначействе на высланные Подольского гражданскою палатого 99
4		1883г. О выдаче жалования чиловникам канцелярли ГанБаря 1883. губернского предвидинеля 24 декабрь 1883г дворянства
5		1896г. Ннига заявлений домовладеньцев г. Каменца 1896г. 3 а 1896г. 1913г фозомумярные списки о службе чиновников да-а 10 казначействе за 1913

Table 2 provides a statistical overview of the collected data, including the mean, variance, and standard deviation for the number of successfully recognized characters, the average complexity rating of texts, and the average accuracy percentage of recognition.

The analysis (Table 2) of the statistical indicators of text recognition results by the OCR model indicates overall effectiveness and challenges associated with processing historical documents. The arithmetic mean of the number of recognized characters is 107.33, with an average complexity rating of text recognition at 2.93, suggesting a moderate level of document complexity. The average percentage of correctly recognized characters is quite high at 71.7%, indicating a reasonably high accuracy of the OCR model. However, significant variance in the number of recognized characters (1908.14) and the average percentage of correctly recognized characters (576.72), as well as the standard deviation for these indicators (43.68 and 24.01, respectively),

underscore the variability in recognition quality among different documents. The correlation value of -0.76 indicates a strong negative relationship between the average complexity rating of texts and the average percentage of correctly recognized characters, demonstrating that as the text complexity increases, the recognition efficiency decreases.

Table 2.
Summary of Expert Evaluation Results

	Characters recognised	SORT	SVPRS
Arithmetic Mean	107,33	2,93	71,7
Variance	1908,14	0,45	576,72
Standard Deviation	43,68	0,67	24,01
Correlation between SORT and SVPRS		-0,76	

The detailed analysis of the data [24] revealed variability in the accuracy of text recognition by the OCR model, which correlates with the complexity rating of the text, ranging from 1.0 to 4.6. Higher SVPRS values, reaching up to 100%, are observed in texts with lower complexity ratings, indicating the high efficiency of the model in recognizing less complex documents. However, a significant decrease in recognition accuracy to 1.74% and below is observed in texts with the highest complexity ratings (4.4 and above), highlighting the limitations of the current OCR model when working with highly complex historical materials. Documents numbered 69-75[24], written in Ukrainian using the Latin alphabet, demonstrated significantly lower recognition accuracy compared to others, as reflected in the average percentage of correctly recognized characters (SVPRS) ranging from 1.74% to 6.99%. Meanwhile, the complexity rating of these texts was the highest among all analyzed documents (4.4 and above), indicating significant challenges that the OCR model faces when working with texts written in the Latin alphabet but in the Ukrainian language. These results underscore the particular challenges associated with recognizing texts in languages using non-standard or less common alphabets and indicate a critical need for the development of specialized approaches and algorithms to enhance OCR accuracy in such conditions.

In conclusion, the analysis of the expert evaluation results and the statistical overview of the collected data regarding the efficiency of the OCR model confirm its significant potential utility in the study of historical texts. However, the high level of negative correlation between text complexity and recognition accuracy emphasizes the importance of further refinement and adaptation of OCR technologies to optimize working with complex historical materials. Special attention to texts written in non-standard alphabets, such as Ukrainian using the Latin alphabet, underscores the necessity for the development of specialized approaches to overcome these unique challenges, paving the way for improving accessibility and preservation of valuable historical heritage.

5. Conclusions

Within this study, an analysis of the effectiveness of the Attention-Gated-CNN-BGRU model for HTR from historical documents stored in the State Archives of Khmelnytskyi Oblast was conducted, primarily focusing on documents written in Ukrainian and Russian languages. The research concentrated on digitized descriptions of ancient deeds from 1861 to 1913. The utilized model demonstrated an SVPRS of 71.7%, indicating significant efficiency in the context of the complexity of the analyzed documents.

Comparing our results with similar solutions [1, 15] in this field, it can be noted that the incorporation of attention mechanism in the Attention-Gated-CNN-BGRU model provided significant advantages in recognition accuracy, especially for texts with high complexity and crossed-out words. This marked a significant advancement compared to traditional CRNN models, which often show decreased efficiency under similar conditions. The identified strong negative correlation (-0.76) between text complexity and recognition accuracy underscores the

necessity for further development and optimization of models for handling highly complex historical manuscripts.

One of the main directions for future research is the development and adaptation of the model for effective recognition of texts written in Ukrainian using the Latin alphabet. This aspect is crucial for the preservation and study of Ukraine's cultural heritage, as a considerable number of historical documents of significant importance are written in this alphabet. The results of our study indicate the potential feasibility of effectively adapting existing technologies for recognizing such texts, but also emphasize the need for further developments in this area.

In conclusion, our research not only confirmed the high effectiveness of the Attention-Gated-CNN-BGRU model in recognizing handwritten text from historical documents but also identified promising directions for future work. Specifically, the development of models for recognizing texts written in Ukrainian using the Latin alphabet could be a key step in preserving and making important historical resources accessible, enriching our understanding of the past and promoting cultural exchange.

6. References

[1] Ukrainian generic handwriting. (б. д.). READ-COOP. https://readcoop.eu/model/ukrainian-generic-handwriting/

[2] Transkribus. (d. b.). Transkribus. https://beta.transkribus.org/sites/tryckfrihet/browse

[3] Dubrovina, L., & Kovalchuk, G. (2016). Development of electronic information resources of manuscript and book heritage at the V. I. Vernadsky National Library of Ukraine. Library Herald, (1), 3-11.

[4] Kruchinina, T. H., & Cherniatynska, Y. H. (2015). Creation of archival collections of websites within the framework of initiative documentation at the Central State Electronic Archives of Ukraine. Archives of Ukraine, (5-6), 61-67.

[5] Maistrenko, A. A., & Romanovsky, R. V. (2018). Digitization of archival documents in foreign countries: information-analytical overview. Archives of Ukraine, (1), 64-87.

[6] Sviatets, Y. A. (2019). Digital sources in research practices of historians: theoretical and methodological aspect. Universum Historiae et Archeologiae= The Universe of History and Archeology. Dnipro, 2(27), 47-68.

[7] Philipova, L. (2018). Digital archives in modern society: terminological and substantive aspects. Library Science. Document Science. Informatics, (2), 6-11.

[8] O. Pisnyi et al., "AR Intelligent Real-time Method for Cultural Heritage Object Recognition," 2023 IEEE 5th International Conference on Advanced Information and Communication Technologies (AICT), Lviv, Ukraine, 2023, pp. 62-66, doi: 10.1109/AICT61584.2023.10452426.

[9] Lipianina-Honcharenko, K., Savchyshyn, R., Sachenko, A., Chaban, A., Kit, I., & Lendiuk, T. (2022). Concept of the Intelligent Guide with AR Support. International Journal of Computing, 271–277. https://doi.org/10.47839/ijc.21.2.2596

[10] Schauer, S., Sieck, J., Lipianina-Honcharenko, K., Sachenko, A., & Kit, I. (2023). Use of Digital Auralised 3D Models of Cultural Heritage Sites for Long-term Preservation. У 2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). IEEE. https://doi.org/10.1109/idaacs58523.2023.10348637

[11] Komar, M., Dorosh, V., Hladiy, G., & Sachenko, A. (2018). Deep Neural Network for Detection of Cyber Attacks. У 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC). IEEE. https://doi.org/10.1109/saic.2018.8516753

[12] Anfilets, S., Bezobrazov, S., Golovko, V., Sachenko, A., Komar, M., Dolny, R., Kasyanik, V., Bykovyy, P., Mikhno, E., & Osolinskyi, O. (2020). Deep multilayer neural network for predicting the winner of football matches. International Journal of Computing, 19(1), 70–77. https://doi.org/10.31891/1727-6209/2020/19/1-70-77

[13] Yatskiv, V., Jun, S., Yatskiv, N., Sachenko, A., & Osolinskiy, O. (2011). Multilevel method of data coding in WSN. У 2011 IEEE 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). IEEE. https://doi.org/10.1109/idaacs.2011.6072894

[14] Sachenko, A., Osolinskyi, O., Bykovyy, P., Dobrowolski, M., & Kochan, V. (2020). Development of the Flexible Traffic Control System Using the LabView and ThingSpeak. У 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT). IEEE. https://doi.org/10.1109/dessert50317.2020.9125036

[15] Aradillas Jaramillo, J. C., Murillo-Fuentes, J. J., & M. Olmos, P. (2018). Boosting Handwriting Text Recognition in Small Databases with Transfer Learning. У 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE. https://doi.org/10.1109/icfhr-2018.2018.00081

[16] Balci, Batuhan, Dan Saadati, and Dan Shiferaw. "Handwritten text recognition using deep learning." CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University, Course Project Report, Spring (2017): 752-759.

[17] Yintong Wang1,2, Wenjie Xiao2 and Shuo Li2 Journal of Physics: Conference Series, Volume 1848, 2021 4th International Conference on Advanced Algorithms and Control Engineering (ICAACE 2021) January 29-31, 2021, Sanya, China

[18] Nisa, H., Thom, J. A., Ciesielski, V., & Tennakoon, R. (2019). A deep learning approach to handwritten text recognition in the presence of struck-out text. У 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE. https://doi.org/10.1109/ivcnz48456.2019.8961024

[19] Yang, J., Ren, P., & Kong, X. (2019). Handwriting Text Recognition Based on Faster R-CNN. У 2019 Chinese Automation Congress (CAC). IEEE. https://doi.org/10.1109/cac48633.2019.8997382

[20] Ingle, R. R., Fujii, Y., Deselaers, T., Baccash, J., & Popat, A. C. (2019). A Scalable Handwritten Text Recognition System. У 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE. https://doi.org/10.1109/icdar.2019.00013

[21] Sanchez, J. A., Romero, V., Toselli, A. H., & Vidal, E. (2016). ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset. У 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE. https://doi.org/10.1109/icfhr.2016.0120

[22] Retsinas, G., Sfikas, G., Gatos, B., & Nikou, C. (2022, May). Best practices for a handwritten text recognition system. In International Workshop on Document Analysis Systems (pp. 247-259). Cham: Springer International Publishing.

[23] Mishra, P., Pai, P., Patel, M., & Sonkusare, R. (2020). Extraction of Information from Handwriting using Optical Character recognition and Neural Networks. У 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE. https://doi.org/10.1109/iceca49313.2020.9297418

[24] GitHub - kniosu/handwriting-text-recognition: Handwriting text recognition. (б. д.). GitHub. https://github.com/kniosu/handwriting-text-recognition

[25] I. Paliy, V. Turchenko, V. Koval, A. Sachenko and G. Markowsky, "Approach to recognition of license plate numbers using neural networks," 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), Budapest, Hungary, 2004, pp. 2965-2970 vol.4, doi: 10.1109/IJCNN.2004.1381137

[26] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, A. Senyk, O. Malanchuk. Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian - Eastern-European Journal of Enterprise Technologies. - 2018. - № 6(2). - P. 19-31.

Vol-3702
urn:nbn:de:0074-3702-6

CMIS 2024
Computer Modeling and Intelligent Systems 2024

Proceedings of The Seventh International Workshop on Computer Modeling and Intelligent Systems (CMIS-2024)

Zaporizhzhia, Ukraine, May 3, 2024.

Edited by

Sergey Subbotin *

* National University "Zaporizhzhia Polytechnic", Faculty of Computer Science and Technologies, Zaporizhzhia, Ukraine

Table of Contents

2024-05-14: submitted by Sergey Subbotin, metadata incl. bibliographic data published under Creative Commons CC0
2024-06-09: published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073) |valid HTML5|

Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data

Khrystyna Lipianina-Honcharenko[1], Mariana Soia[1], Khrystyna Yurkiv[1], Andrii Ivasechko[1].
[1] *West Ukrainian National University, Lvivska str., 11, Ternopil, 46000, Ukraine*

Abstract
In today's world, where information spreads with unprecedented speed, disinformation poses a serious challenge to public trust and information security. The full-scale invasion of Ukraine by Russia in 2022 activated the use of disinformation as a tool of hybrid warfare, highlighting the need for effective methods of identification and control. This article focuses on evaluating the effectiveness of various machine learning methods for detecting disinformation in Ukrainian text data, using a dataset that includes news headlines collected during the conflict. The study encompasses the analysis of logistic regression, support vector machines (SVM), random forest, gradient boosting, KNN, decision trees, XGBoost, and AdaBoost. Model evaluation was performed using standard metrics: precision, recall, F1-score, overall accuracy, and confusion matrix. The results indicate significant potential for using machine learning in the fight against disinformation, particularly the random forest model demonstrated the highest effectiveness. The study emphasizes the importance of adapting and optimizing classifiers for the specific task of disinformation analysis, paving the way for further research in this field.

Keywords
Disinformation, machine learning, classification, text data.

1. Introduction

In the modern information space, a vast amount of data is generated daily, a significant portion of which is news content. In the context of increasing globalization and accessibility of information technologies, information spreads rapidly through the network, making it a powerful tool for influencing public opinion. However, this also paves the way for the mass dissemination of disinformation, which can have significant consequences for society, politics, and international relations. Navigating this flow of information and distinguishing reliable data from false has become an increasingly important task.

The war that began with Russia's full-scale invasion of Ukraine in February 2022 is a striking example of the use of disinformation as a weapon in hybrid warfare. This has created a need for the development of effective tools for analyzing and classifying informational content, with the goal of identifying and counteracting disinformation.

In this context, machine learning and natural language processing methods play a key role in the detection and analysis of fake news. The application of these technologies allows for the automation of the disinformation detection process, providing fast and efficient processing of large volumes of data. At the same time, the development of effective machine learning models for information classification requires a deep understanding of data specifics, preprocessing methods, and model optimization.

This article is devoted to the analysis of a dataset containing news headlines collected during the Russo-Ukrainian conflict, aimed at identifying disinformation. It considers the application of various machine learning methods, including logistic regression, support vector machines (SVM), random forest, gradient boosting, KNN, decision trees, XGBoost, and AdaBoost for the classification of text data. The concluding section is dedicated to evaluating the effectiveness of these models using standard evaluation metrics, including precision, recall, F1-score, overall

Proceedings Acronym: Proceedings Name, Month XX–XX, YYYY, City, Country
xrustya.com@gmail.com (K. Lipianina-Honcharenko); mariankasoa@gmail.com (M. Soia); kh.yurkiv@wunu.edu.ua (K. Yurkiv); Andrewivasechko@gmail.com (A. Ivasechko).
0000-0002-2441-6292 (K. Lipianina-Honcharenko); 0009-0001-3719-6460 (M.Soia); 0009-0007-4917-3251 (K. Yurkiv); 0009-0002-8623-7828 (A. Ivasechko).

accuracy, and the confusion matrix, which allows determining the most effective methods for combating disinformation in the context of information warfare.

2. Related Work

In contemporary research in the field of information security and media content analysis, the importance of detecting and analyzing disinformation, especially anti-vaccine content on social media platforms such as Twitter, has gained particular relevance. In [1], language-neutral models are developed for detecting such content on a large scale, using multifaceted representations of messages in networks. Meanwhile, [2] focuses on the challenges associated with pre-training graph neural networks for context-oriented detection of fake news, pointing out strategic and resource constraints. In [3], a comparative analysis of supervised and unsupervised machine learning algorithms for detecting fake news is conducted, demonstrating their performance, efficiency, and robustness.

Research covering the analysis of disinformation and public opinion on the Russo-Ukrainian War employs a variety of methodologies, including sentiment analysis, creation and analysis of datasets, and studies of the impact of war on language choice. One study models and clusters sentiment trends of different countries regarding the war [4], while another offers a detailed dataset of tweets related to the crisis [5]. The discourse on Twitter about the war is also analyzed, with a particular focus on language and the geographical origin of tweets [6]. Another study focuses on the challenges of labeling sensitive content and its psychological impact on annotators [7]. It is also examined how the war affects the language choice of Ukrainians on Twitter, analyzing changes in language preferences over time [8]. A separate study provides insights into the activity on subreddits related to the conflict, analyzing post volumes, comments, and the level of engagement [9]. Research [10] demonstrates that the application of the BERT model for fake news detection achieves an accuracy of 79.88% and an area under the ROC curve of 0.87, highlighting its potential in combating disinformation on social networks.

As confirmed by the aforementioned analysis, research in the field of disinformation detection, particularly fake news, is a significant direction in the context of information security and media content analysis. The need for the development and application of advanced technological solutions for effective fake news detection is particularly compelling. However, there is a noticeable lack of research focused on the analysis of disinformation in the Ukrainian language, which poses a challenge to the scientific community to expand the linguistic spectrum of research in this field. Considering this, the aim of this study is to develop intelligent methods for detecting disinformation, specifically fake news, with an emphasis on Ukrainian-language content. This will enhance the level of information security and ensure the integrity of the news space in the conditions of the modern information society.

3. Research methodology

3.1 Dataset Description

For data collection and processing, the dataset (Ukrainian language) [11] was used, which contains approximately 10,700 news headlines about the Russo-Ukrainian War, collected from February 24 to December 11, 2022, covering the period from the beginning of the full-scale invasion.

The dataset for the analysis of disinformation in the Ukrainian media space consists of two main files: "data_set_4.csv" (Table 1) and "news_data.csv" (Table 2), each containing news headlines classified as true ("True") or false ("False"). The "data_set_4.csv" file records 8,237 true and 2,498 false news items, while "news_data.csv" contains a significantly larger number of true news items — 48,006, compared to 2,024 false, reflecting a wide range of informational content collected for the study of the dissemination of disinformation during the Russo-Ukrainian War.

Both datasets (see Table 1,2) are used for the analysis of disinformation in the Ukrainian media space and include data that were collected from official and unofficial sources with the purpose of studying the spread of fake news in the context of the Russo-Ukrainian War.

Table 1
Structure of data_set_4.csv

Field	Description	Data Type
Text	News Headline	Text
Label	Label that marks the news as true ("True") or false ("False")	Boolean
Link	Link to the news source (available only in this file)	Text

Table 2
Structure of news_data.csv

Field	Description	Data Type
Text	News Headline	Text
Label	Classification of the news as true ("True") or false ("False")	Boolean

The graphs (Fig.1) of label distribution show that in both datasets, the number of true messages predominates over the false ones. For example, in the first dataset, the ratio of true messages to false is high, which indicates a focus on reliable information.

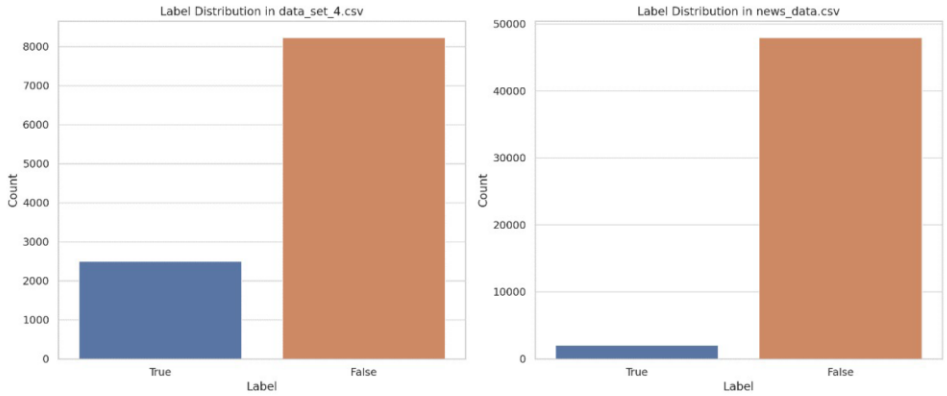


Figure 1: Label Distribution

The analysis of the most frequently used words (Fig.2) in the first dataset revealed words that appear hundreds of times. This allows identifying key topics of discussions or news, for example, the word "Ukraine" may occur most frequently, emphasizing the geographical or political focus of the collected data.

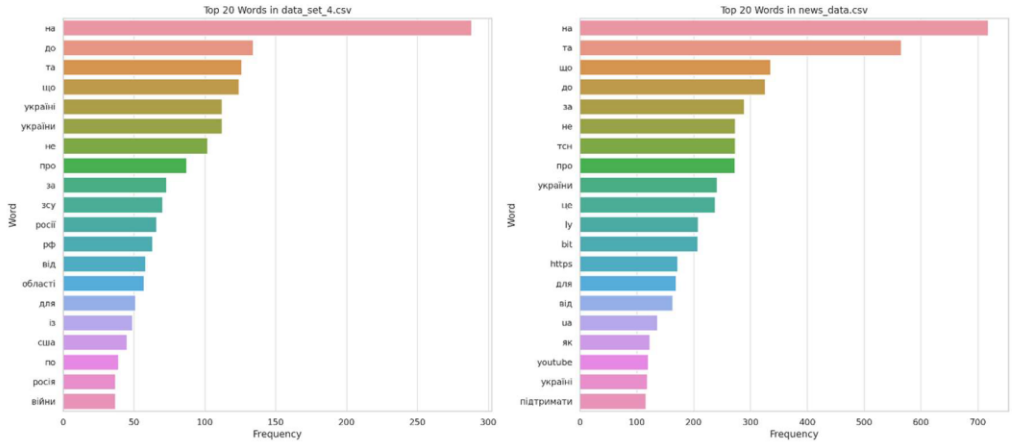


Figure 2: Top 20 Words

Most texts (Fig.3) in the first dataset have a length of 100 to 500 characters. Such information helps to understand the average volume of messages, which may indicate a prevalence of short news or overviews.

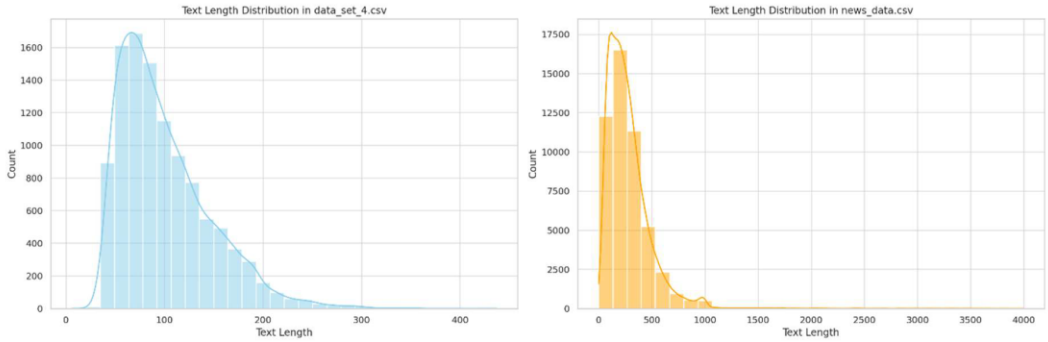


Figure 3: Distribution of Text Lengths

Analysis of the presented datasets covering headlines of news about the Russian-Ukrainian war demonstrates a significant advantage of credible information compared to false, emphasizing the focus on quality content. Considering that the dataset "news_data.csv" contains substantially more records compared to "data_set_4.csv", it is logical to use the former for training machine learning models as it provides a broader range of information and a greater number of examples for training. Meanwhile, the smaller dataset "data_set_4.csv" can serve as an excellent set for testing and evaluating the effectiveness of models on a smaller but specific data sample. This will allow assessing the model's ability to generalize learning on new, previously unknown data, emphasizing its practical value in real-world disinformation analysis conditions.

3.2 Description of used classifiers

In modern data analysis, especially in the context of detecting misinformation, the use of machine learning algorithms for classifying textual data becomes a key tool for developers and analysts. The diversity of classification methods [12], such as logistic regression, support vector machines (SVM), random forest, gradient boosting, k-nearest neighbors (KNN), decision tree, XGBoost, and AdaBoost, provides a wide range of approaches for data analysis and classification. Each of these algorithms has its unique advantages and limitations, making them more or less suitable for specific types of data and analytical tasks. The main goal is to select the optimal classifier that best fits the specificity of the task and the data we are working with.

Logistic regression [13] is a classification method used to predict the probability of two possible outcomes based on one or more independent variables. It transforms the linear combination of input data into probability using the logistic function. The advantages of logistic regression include simplicity in interpreting results, but it may be limited when analyzing complex relationships between variables and requires an assumption of linearity in relationships. Mathematically, logistic regression models the probability $P(Y=1)$ as a function of X, where Y is the dependent variable, and X is the set of independent variables. The probability is described by the equation:

$$P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_k X_k)}}, \quad (1)$$

where e is the base of the natural logarithm, β_0 is the constant term (intercept), and $\beta_1,\ldots,\beta_k$ are the coefficients of the independent variables. This equation allows us to estimate the probability that an observation belongs to class 1, depending on the values of the independent variables.

Support Vector Machine (SVM) algorithm [14] finds a hyperplane in a multidimensional space that best separates different data classes, maximizing the distance between the closest data points (support vectors) of different classes. The advantages of SVM include high accuracy in classification tasks, especially on relatively small datasets, and flexibility through the use of various kernel functions. However, its drawbacks include high computational resource

requirements for large datasets and complexity in interpreting the model. Mathematically, SVM seeks to solve the optimization problem: minimize $\frac{1}{2} || w ||^2$ subject to the constraints that

$$y_i (w \cdot x_i + b) \geq 1 \text{ to all } I, \quad (2)$$

where w is the weight vector of the hyperplane, b is the bias, x_i are the feature vectors, and y_i are the class labels.

The Random Forest algorithm [15] creates an ensemble of decision trees, training each tree on randomly selected subsets of the training dataset and features, which ensures high accuracy and model universality. The advantages of Random Forest include its ability to efficiently handle large datasets with high feature dimensionality and its lower tendency to overfit compared to individual decision trees. However, its drawbacks include relatively high computational resource requirements and the complexity of interpreting the model due to the large number of trees. The mathematical interpretation of Random Forest is based on the principle of "wisdom of the crowd," where the final model decision is determined by voting among the trees for classification tasks or averaging the outputs for regression tasks. More precisely, for classification:

$$Y = mode\{y_1, y_2, \ldots, y_n\}, \quad (3)$$

and for regression:

$$Y = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad (4)$$

where y_i is the prediction of each tree, and Y is the final prediction of the ensemble.

Gradient Boosting [16] is an ensemble machine learning method that improves predictions by sequentially training weak models, typically decision trees, to minimize a loss function. It is characterized by high prediction accuracy and flexibility in parameter tuning, but it can be prone to overfitting if not properly configured and requires more time and computational resources for training compared to other algorithms. Mathematically, Gradient Boosting performs optimization by adaptively reducing the difference between actual and predicted values using gradient descent, where the model update in the m-th iteration is defined as:

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x), \quad (5)$$

where $F_{m-1}(x)$ is the prediction at the previous step, $h_m(x)$ is the weak classifier, and α_m is the learning rate.

The k-Nearest Neighbors (KNN) algorithm [17] classifies objects based on the nearest training examples in the feature space, where "k" indicates the number of neighbors considered to determine the class of a new object. The advantages of KNN include ease of implementation and its ability to effectively work with multi-class datasets. However, it requires significant computational resources to store training data and determine neighbors in large datasets, and it is sensitive to irrelevant features and data scaling. Mathematically, the classification of an object x in the KNN algorithm is determined by the majority vote of its neighbors, where each neighbor is weighted according to the inverse distance to x, typically using Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{ij})^2}, \quad (6)$$

where x is the point for classification, x_i is a point from the training dataset, and j varies from 1 to m, the number of features. The class of x is determined based on the most frequent class among the k nearest neighbors.

The Decision Tree algorithm [18] builds a predictive model in the form of a tree-like structure by partitioning the dataset into smaller subsets while simultaneously developing the associated decision tree. The advantages of decision trees include ease of interpretation, the ability to handle both numerical and categorical data, and no requirement for data normalization. However, decision trees are prone to overfitting, especially with deep trees, and can be unstable, meaning small changes in data can result in significantly different decision trees. Mathematically, decision trees use the concept of information gain or reduction in uncertainty (entropy) to select the attribute that best splits the dataset into subsets according to the target variable. The information gain for attribute A is calculated as:

$$IG(T, A) = Entropy(T) - \sum_{v \in Values(A)} \frac{|T_v|}{|T|} Entropy(T_v), \quad (7)$$

where T is the training set, $Values(A)$ is the set of all possible values of attribute A, T_v is the subset of T for which attribute A has the value v, and $Entropy(T)$ is the entropy of the training set T.

XGBoost (eXtreme Gradient Boosting) [19] is a highly efficient implementation of the gradient boosting algorithm, which optimizes both linear models and decision trees. The algorithm stands out for its high execution speed, ability to efficiently scale to large datasets, and built-in support for regularization, helping to mitigate overfitting. However, XGBoost can be challenging to tune due to the large number of hyperparameters and requires more time for training compared to simpler models. Mathematically, XGBoost minimizes losses using gradient descent, where the objective function includes both the loss function L and a penalty for model complexity Ω, adding regularization:

$$Obj = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad (8)$$

where y_i are the true labels, $\hat{y}_i$ are the predicted labels, f_k are the functions representing individual trees, and $\Omega(f_k)$ includes terms for regularization, such as the number of leaves and the sum of squares of node weights, thereby reducing the risk of overfitting.

AdaBoost (Adaptive Boosting) [20] is a machine learning algorithm that combines multiple weak classifiers to create a strong classifier, using an iterative approach to correct the errors of previous classifiers by assigning greater weight to observations that are harder to classify. The advantage of AdaBoost is its ability to improve prediction accuracy, ease of implementation, and automatic correction of underperforming classifiers. However, it can be prone to overfitting in the presence of outliers or highly noisy data and requires careful tuning of the number of iterations. Mathematically, AdaBoost adapts the weights of training observations, w_i, by increasing the weights of incorrectly classified observations. At each iteration step t, a classifier h_t is selected to minimize the weighted sum of errors. The final classifier is determined as a weighted sum of these classifiers.

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)), \quad (9)$$

where α_t is the weight assigned to classifier h_t, which depends on its accuracy.

Conclusions drawn from the description of the used classifiers underscore the importance of adapting the model to the specificity of the dataset and the analytical task. The effectiveness of each method depends on the size and quality of the data, the complexity of relationships in the dataset, as well as specific requirements for accuracy and interpretation of results. In the context of disinformation analysis, the choice between the simplicity of interpreting logistic regression and the high accuracy but complexity of tuning XGBoost or SVM may determine the success or failure in identifying fake news. Thus, careful selection and tuning of classifiers are critically important for developing effective tools to combat disinformation in the context of the Russian-Ukrainian war.

3.3 Evaluation metrics

For evaluating the effectiveness of models in classification tasks, key metrics such as precision, recall, F1-score, accuracy, and confusion matrix are utilized [21]. These metrics allow for a deeper analysis of the model's performance, identifying potential weaknesses, and optimizing the model for better results.

Precision is defined as the ratio of the number of true positive results to the total number of results classified as positive by the model. Mathematically, it is represented as:

$$P = \frac{TP}{TP+FP}, \quad (10)$$

where TP — true positives, and FP — false positives.

Recall measures the model's ability to identify all actual positive cases in the dataset. It is defined as the ratio of the number of correctly identified positive results to the sum of correctly identified positive results and instances that are actually positive but were missed by the model. The formula for calculation is:

$$R = \frac{TP}{TP+FN}, \quad (11)$$

where FN — false negatives.

The F1 Score is the harmonic mean between precision and recall, providing a balance between these two metrics. This is particularly useful in situations where class imbalances may cause biases in one metric over the other. F1 is defined as:

$$F1 = 2 \cdot \frac{P \cdot R}{P+R}. \quad (12)$$

Accuracy measures the percentage of cases correctly classified by the model and is defined by the formula:

$$A = \frac{TP+TN}{TP+TN+FP+F}, \quad (13)$$

where TN — true negatives.

The Confusion Matrix provides a visualization of classification results by representing the counts of TP, TN, FP, and FN in the form of a matrix. This allows us not only to determine the accuracy of the model but also to understand the types of errors made by the model.

3.4 Model training

The training procedure (Figure 4) on the training dataset is a fundamental step in the development of effective machine learning algorithms for text data classification. In this context, the use of datasets "news_data.csv" and "data_set_4.csv" for training and testing models, respectively, provides a valuable foundation for misinformation analysis. The initialization of the procedure begins with the import and preprocessing of data, including the removal of records without textual content, ensuring data cleanliness for subsequent processing stages.

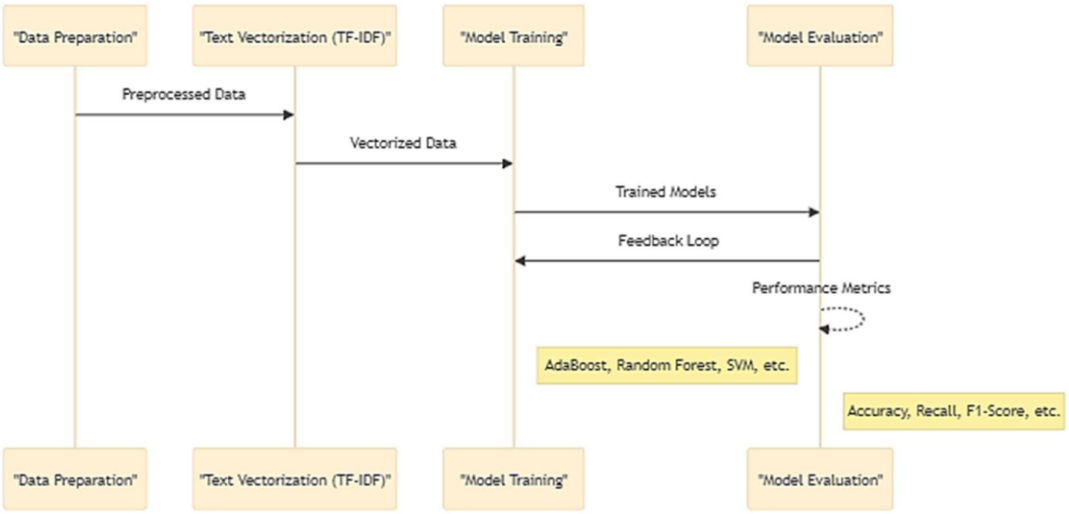


Figure 4: Model Training Process

Text vectorization using TF-IDF (Term Frequency-Inverse Document Frequency) is a key step in data preparation, as it transforms textual data into a numerical format, making them suitable for processing by machine learning models. This method considers not only the frequency of words in the text but also their uniqueness through the inverse frequency of documents, allowing the model to better identify important features in the text.

After preparing the vectorized training and testing data, the next stage involves training different models on the training dataset. This process requires the application of machine learning algorithms to the training dataset to form a model capable of effectively classifying text based on learned features. Each model adjusts its parameters to minimize errors on the training set while simultaneously ensuring the ability to generalize to new data.

Evaluation of the effectiveness of each trained model on the test dataset is crucial for determining its suitability and efficiency in classification tasks. The use of evaluation metrics such as accuracy, recall, F1-score, and overall accuracy allows for deep analysis and comparison of model performance. The evaluation results can be visualized for better understanding of model effectiveness, and the best-performing model can be selected for further use or saved for future analysis.

Thus, the model training procedure on the training dataset using pre-processed and vectorized text data is fundamental for developing reliable machine learning tools capable of effectively classifying and analyzing text for misinformation.

4. Research results

Further, we will discuss the implementation and evaluation of the effectiveness of various machine learning models in the task of classifying misinformation data in the Ukrainian media space, contextualized in the context of Russia's full-scale intervention. By analyzing a wide range of algorithms, we aim to identify the most effective methods for accurate detection and differentiation of misinformation incidents. This analysis is based on a comprehensive comparison of overall classification accuracy as well as specific metrics such as precision, recall, and F1-score for each class. The implementation was done in the Python programming language utilizing libraries for text analysis.

The logistic regression model showed (Figure 5) an overall classification accuracy of 86.4% on the test sample of 10,735 examples, demonstrating high effectiveness in identifying true values with an F1-score of 0.92 for the "True" class. However, the model was less accurate in identifying the "False" class, with a prediction accuracy of 0.96 and a low recall score of 0.44, indicating a significant number of false negatives, as reflected in the confusion matrix with 1,407 misclassified examples.

Model: logistic_regression

	precision	recall	f1-score	support
False	0.96	0.44	0.60	2498
True	0.85	0.99	0.92	8237
accuracy			0.86	10735
macro avg	0.90	0.72	0.76	10735
weighted avg	0.88	0.86	0.84	10735

Accuracy: 0.864182580344667

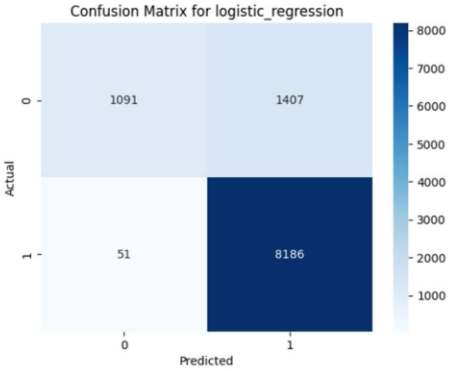


Figure 5: LogisticRegression Results

The SVM model demonstrated (Figure 6) high overall classification accuracy of 93.6% for the test sample, indicating its effectiveness in distinguishing between the "True" and "False" classes. Specific accuracy and recall metrics for the "False" class were 0.97 and 0.75, respectively, demonstrating the model's ability to identify negative cases well, albeit with some errors. At the same time, high metrics for the "True" class with precision of 0.93 and recall of 0.99 indicate minimal false negative classifications, confirmed by a low number of errors in the confusion matrix (618 for "False" and 68 for "True").

Model: svm

	precision	recall	f1-score	support
False	0.97	0.75	0.85	2498
True	0.93	0.99	0.96	8237
accuracy			0.94	10735
macro avg	0.95	0.87	0.90	10735
weighted avg	0.94	0.94	0.93	10735

Accuracy: 0.9360968793665579

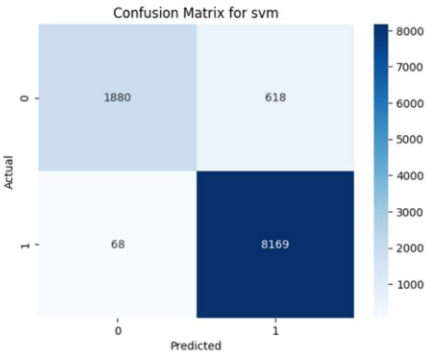


Figure 6: SVM Results

The Random Forest model demonstrated (Figure 7) high accuracy in classification with an overall accuracy of 95.3% on the test dataset, indicating the model's high effectiveness in recognizing both classes. For the "False" class, the model showed high accuracy (0.98) and a relatively high recall of 0.81, indicating the model's ability to effectively identify negative cases with a moderate number of errors. Conversely, extremely high accuracy (0.95) and almost perfect

recall (1.00) for the "True" class highlight the minimal number of false negative results, corroborated by the low number of errors in the confusion matrix.

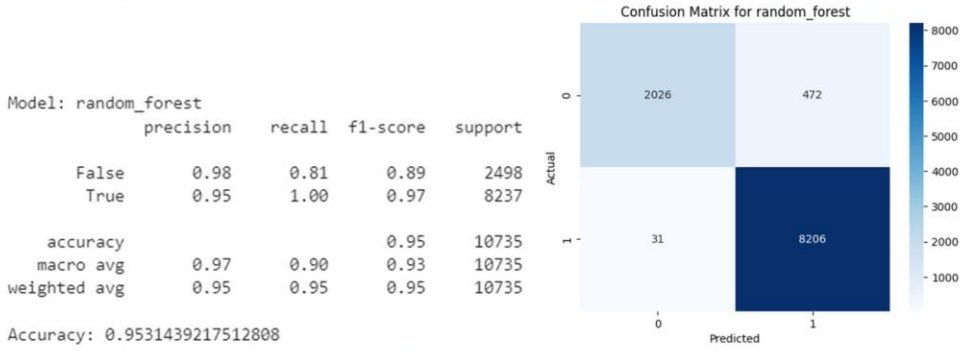


Figure 7: RandomForestClassifier Results

The Gradient Boosting model achieved (Figure 8) an overall classification accuracy of 87.3% on the test dataset, indicating the model's good ability to distinguish between the "True" and "False" classes. The model's precision for the "False" class was high (0.94), but the recall was only 0.48, indicating a significant number of type II errors, where negative cases are often misclassified as positive. Conversely, for the "True" class, the model showed impressive classification ability with a precision of 0.86 and a recall of 0.99, demonstrating its high effectiveness in identifying true positive cases with minimal errors.

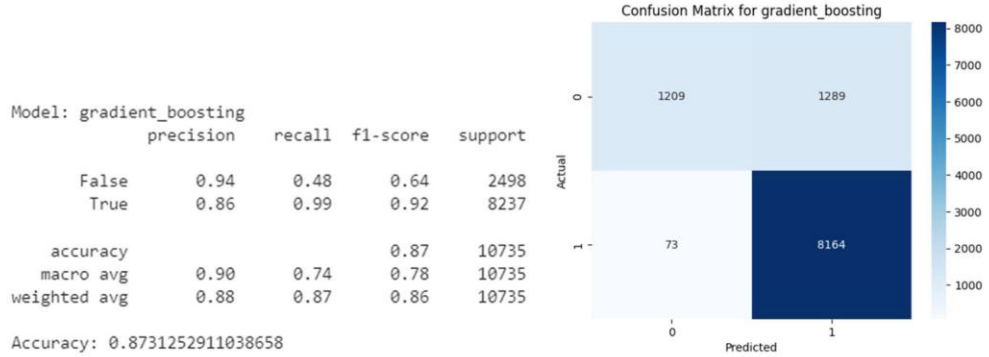


Figure 8: Gradient Boosting Results

The KNN (K-Nearest Neighbors) model demonstrated (Figure 9) an overall classification accuracy of 87.6% on the test dataset, highlighting its ability to effectively classify data. Although the model showed high precision (0.91) for the "False" class, the recall was only 0.51, indicating difficulties in identifying all negative cases. Conversely, for the "True" class, the model exhibited excellent precision (0.87) and a high recall (0.99), indicating its ability to identify positive cases with high confidence, albeit with some false positive errors, as seen in the confusion matrix.

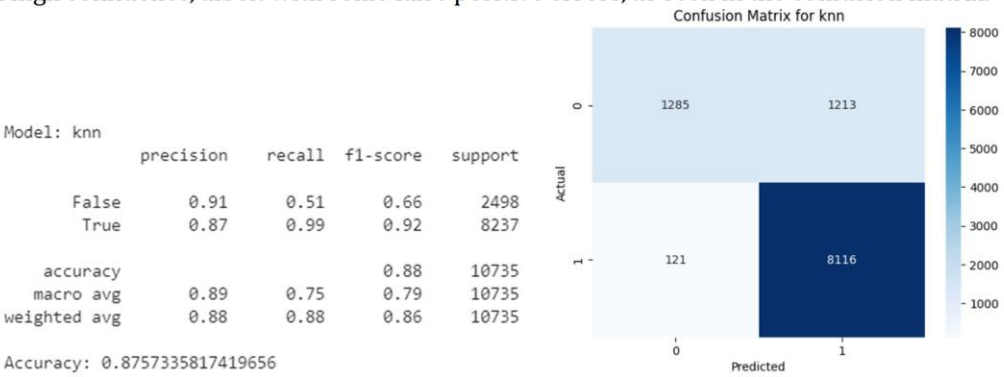


Figure 9: KNN Results

The Decision Tree model achieved (Figure 10) a high level of classification accuracy, 91.4%, on the test dataset, confirming its effectiveness in classifying data into "True" and "False" classes. For the "False" class, the model showed relatively high precision (0.81) and recall (0.83), indicating a balanced ability to identify negative cases with a relatively small number of errors. Meanwhile, for the "True" class, the model provided impressive precision (0.95) and recall (0.94), demonstrating its strong capabilities in identifying positive cases with minimal false negatives, as reflected in the confusion matrix.

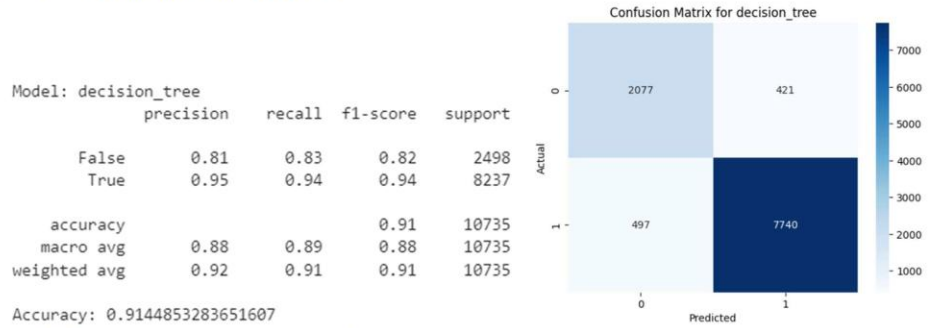


Figure 10: DecisionTreeClassifier Results

The XGBoost model demonstrated (Figure 11) an overall accuracy of 89.2% on the test dataset, indicating its ability to effectively classify the given dataset. The precision and recall metrics for the "False" class were 0.89 and 0.61, respectively, indicating the model's higher ability to correctly identify negative cases, albeit with some errors. Meanwhile, for the "True" class, the model showed high precision and recall (both 0.89 and 0.98), demonstrating excellent ability to accurately identify positive cases with minimal errors, as reflected in its high F1-score.

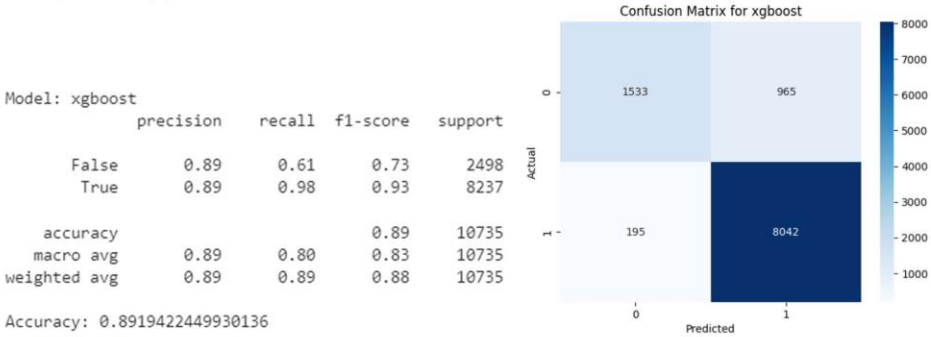


Figure 11: XGBoost Results

The AdaBoost model exhibited (Figure 12) an overall classification accuracy of 86.9% on the test dataset, indicating its effectiveness in recognizing data, albeit with some limitations. For the "False" class, the model achieved a precision of 0.88 with a recall of 0.50, indicating a relatively low ability to identify all negative cases. On the other hand, high precision (0.87) and recall (0.98) for the "True" class underscore the model's strong ability to detect positive cases, albeit with few errors, as reflected in the confusion matrix.

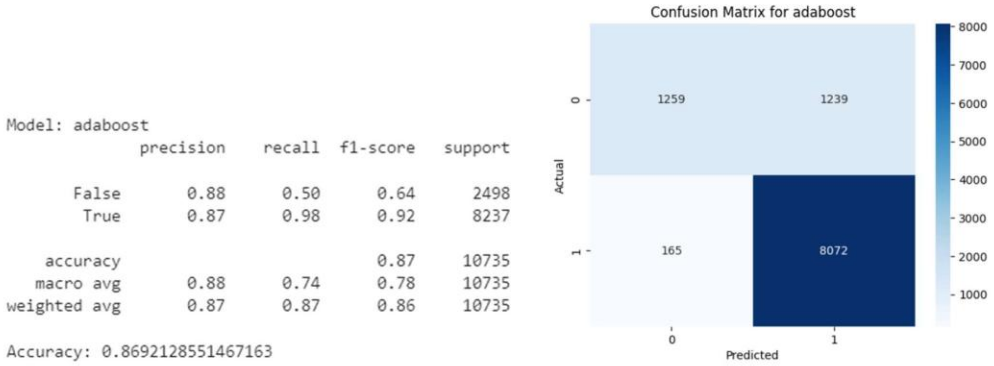


Figure 12: AdaBoostClassifier Results

Therefore, analyzing the results of applying various machine learning models to classify disinformation data in the Ukrainian media space after the onset of Russia's full-scale invasion, it can be noted that the Random Forest model proved to be the most effective with an accuracy of 95.3%, highlighting its ability to accurately detect and differentiate disinformation incidents. At the same time, models such as AdaBoost and logistic regression showed lower overall accuracy, which may indicate their limitations in identifying subtle cases of disinformation or more subjective aspects of information operations. This underscores the importance of choosing the appropriate model for the task of analyzing disinformation, where Random Forest may be more suitable for deep understanding and detecting complex patterns of disinformation in the context of information warfare.

5. Conclusion

The analysis of the effectiveness of various machine learning models applied to the task of classifying news headlines for misinformation in the Ukrainian media space demonstrates significant variations in accuracy, recall, and F1-score among the models. The considered algorithms, including logistic regression, SVM, random forest, gradient boosting, KNN, decision tree, XGBoost, and AdaBoost, showed different levels of performance in addressing the task.

The random forest model emerged as the most effective, achieving an overall accuracy of 95.3%, indicating its high capability to recognize and distinguish true and false messages. This is supported by high precision scores for both the "False" class (0.98) and the "True" class (0.95), as well as significant recall scores for both classes (0.81 for "False" and 1.00 for "True"), demonstrating its effectiveness in minimizing both false positives and false negatives.

These results underscore the importance of choosing the appropriate model for a specific misinformation analysis task. The model choice not only affects the overall classification accuracy but also the model's ability to minimize false positives or false negatives, which is crucial for developing effective tools to combat misinformation. Particularly, the random forest model, which demonstrated the best performance, can be recommended as the optimal choice for similar tasks, providing a high level of accuracy and the ability to effectively distinguish between true and false messages.

Further scientific research in the field of identification and analysis of misinformation in the Ukrainian media space requires deeper development and improvement of machine learning algorithms, with a particular focus on enhancing their ability to recognize subtle and complex forms of misinformation. The results of our study show that the random forest model, with an accuracy of 95.3%, proved to be the most effective, but there is potential for improvement, especially in accurately distinguishing between true and false messages. In the future, researchers may focus on developing hybrid models that combine the advantages of multiple algorithms, including deep learning and neural networks, to ensure greater adaptability and accuracy in different informational contexts. Additionally, an important direction will be the development of methods that allow models to better understand the semantic context and emotional tone of texts, which can significantly improve their ability to identify hidden

misinformation. Implementing such approaches will require not only technological innovations but also a deeper understanding of linguistic nuances and cultural-historical contexts on which misinformation is based.

References

[1] Ashford, J.R. (2024). Detecting Anti-vaccine Content on Twitter using Multiple Message-Based Network Representations. arXiv preprint arXiv:2402.18335.

[2] Donabauer, G., & Kruschwitz, U. (2024). Challenges in Pre-Training Graph Neural Networks for Context-Based Fake News Detection: An Evaluation of Current Strategies and Resource Limitations. arXiv preprint arXiv:2402.18179.

[3] Kaur, S., & Ranjan, S. (2024). Comparative Analysis of Supervised and Unsupervised Machine Learning Algorithms for Fake News Detection: Performance, Efficiency, and Robustness. ResearchGate.

[4] Vahdat-Nejad, H., Akbari, M. G., Salmani, F., Azizi, F., & Nili-Sani, H. R. (2023). Russia-Ukraine war: Modeling and Clustering the Sentiments Trends of Various Countries. arXiv preprint arXiv:2301.00604.

[5] Haq, E. U., Tyson, G., Lee, L. H., Braud, T., & Hui, P. (2022). Twitter dataset for 2022 russo-ukrainian crisis. arXiv preprint arXiv:2203.02955.

[6] Zia, H. B., Haq, E. U., Castro, I., Hui, P., & Tyson, G. (2023). An Analysis of Twitter Discourse on the War Between Russia and Ukraine. arXiv preprint arXiv:2306.11390.

[7] Daria, S. (2023). When a Language Question Is at Stake. A Revisited Approach to Label Sensitive Content. arXiv preprint arXiv:2311.10514.

[8] Racek, D., Davidson, B. I., Thurner, P. W., & Kauermann, G. (2023). The Politics of Language Choice: How the Russian-Ukrainian War Influences Ukrainians' Language Use on Twitter. arXiv preprint arXiv:2305.02770.

[9] Zhu, Y., Haq, E. U., Lee, L. H., Tyson, G., & Hui, P. (2022). A reddit dataset for the russo-ukrainian conflict in 2022. arXiv preprint arXiv:2206.05107.

[10] Padalko, H., Chomko, V., & Chumachenko, D. (2023). Misinformation Detection in Political News using BERT Model. Proceedings of the 3rd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2023) Waterloo, Canada, November 20-22, 2023. 117-127

[11] Ukrainian news. (n.d.-a). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news/data?select=news_data.csv

[12] Lipyanina, H., Maksymovych, V., Sachenko, A., Lendyuk, T., Fomenko, A., & Kit, I. (2020, August). Assessing the investment risk of virtual IT company based on machine learning. In International Conference on Data Stream Mining and Processing (pp. 167-187). Cham: Springer International Publishing.

[13] Lipyanina, H., Sachenko, S., Lendyuk, T., Brych, V., Yatskiv, V., & Osolinskiy, O. (2021, January). Method of detecting a fictitious company on the machine learning base. In International Conference on Computer Science, Engineering and Education Applications (pp. 138-146). Cham: Springer International Publishing.

[14] Kalogridis, I. (2024). Robust and adaptive functional logistic regression. Computational Statistics & Data Analysis, 192, 107905.

[15] Çakir, M., Yilmaz, M., Oral, M. A., Kazanci, H. Ö., & Oral, O. (2023). Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. Journal of King Saud University-Science, 35(6), 102754.

[16] Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

[17] Wang, C., Xiao, W., & Liu, J. (2023). Developing an improved extreme gradient boosting model for predicting the international roughness index of rigid pavement. Construction and Building Materials, 408, 133523.

[18] Çakir, M., Yilmaz, M., Oral, M. A., Kazanci, H. Ö., & Oral, O. (2023). Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. Journal of King Saud University-Science, 35(6), 102754.
[19] Gao, B., Zhou, Q., & Deng, Y. (2024). HIE-EDT: Hierarchical interval estimation-based evidential decision tree. Pattern Recognition, 146, 110040.
[20] Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environmental Modelling & Software, 105971.
[21] Xing, H. J., Liu, W. T., & Wang, X. Z. (2024). Bounded exponential loss function based AdaBoost ensemble of OCSVMs. Pattern Recognition, 148, 110191.

Rule-based method for aligning media content with ukrainian legislation

Khrystyna Lipianina-Honcharenko [1,†], Ihor Ihnatiev[1,†], Hennadii Bohuta [1,†], Khrystyna Yurkiv[1,*,†] and Stanislav Novosad [1,†]

[1] West Ukrainian National University, Lvivska str., 11, Ternopil, 46009, Ukraine

Abstract

We introduce a novel rule-based system for automatically mapping unstructured media texts to specific articles of Ukrainian legislation, addressing the growing need for transparent, explainable tools in legal and security monitoring. Leveraging expert-crafted lexical dictionaries for twelve regulatory norms and a threshold-based matching algorithm, our method achieves balanced performance (Precision – 0.84, Recall – 0.83, F1 – 0.84) on a diverse test set of news articles. The system is delivered as an interactive Streamlit application that supports dynamic dictionary updates and simultaneous sentiment analysis, enabling users to assess both legal relevance and emotional tone of content. Through 15 real-world case studies, we demonstrate the approach's practical utility in governmental and media-watch contexts and discuss paths for expanding dictionary coverage and lowering detection thresholds for shorter texts. Our work extends prior research on rule-based text analysis in domains such as cybersecurity and social media, and contributes a reproducible Explainable AI framework tailored for Ukrainian legal monitoring.

Keywords

rule-based classification; legal text analysis; media monitoring; Ukrainian legislation; explainable AI.

1. Introduction

In the context of the previous study [1], which implemented a thematic text classification system based on emotional coloring with an achieved accuracy of 92%, this paper proposes an improved approach to a related, yet significantly narrower task – the automatic matching of input text to one of the predefined articles of Ukrainian legislation. Unlike general categorization aimed at broad semantic classification, the proposed method enables the establishment of a direct normative link between the content and a specific legal provision.

In the current context of hybrid threats and information attacks, the relevance of automated legal monitoring systems is significantly increasing, especially in terms of ensuring the internal security of the state. Previous research in the field of cybersecurity has demonstrated the effectiveness of neural networks in detecting attacks [2] and in building intelligent cyber defense systems based on artificial immunity models [3], which confirms the potential of specialized rule-based approaches in the tasks of protecting the information space. At the same time, decision support models in the field of internal security [4] and modeling user responses on social networks [5] indicate a growing need for transparent and reproducible algorithms for analyzing sensitive content. The system we propose for formalized alignment of media texts with legal norms is a logical extension of these approaches and has the potential to be integrated into the information security architecture at the state or institutional monitoring level.

The methodology is based on a rule-based approach that utilizes expert-defined dictionaries of key terms for each of the 12 legal articles. By applying a phrase occurrence counter and threshold filtering, the system enables highly interpretable identification of the most relevant article or confirmation of its absence. This approach is particularly valuable in the context of automated

MoMLeT-2025: 7th International Workshop on Modern Machine Learning Technologies, June, 14, 2025, Lviv-Shatsk, Ukraine

*Corresponding author.
†These authors contributed equally.
xrustya.com@gmail.com (Kh. Lipianina-Honcharenko); iiv@wunu.edu.ua (I. Ihnatiev); genabogata7@gmail.com (H. Bohuta); kh.yurkiv@wunu.edu.ua (Kh. Yurkiv), stasnovosad1998@gmail.com (St. Novosad)
0000-0002-2441-6292 (K. Lipianina-Honcharenko); 0000-0003-0729-9247 (I. Ihnatiev); 0009-0000-9788-1753 (H. Bohuta); 0009-0007-4917-3251 (Kh. Yurkiv), 0000-0002-0404-1147 (St. Novosad)

systems for preliminary legal analysis, such as in the monitoring of news, social media, citizen appeals, or expert commentary.

The study also implements a sentiment analysis component based on lexicon-oriented evaluation, which allows for simultaneous assessment of both normative relevance and the overall emotional tone of a message. The integration of these two dimensions — normative and tonal — enables the development of an Explainable AI tool for legal monitoring that is both transparent and applicable without the need for machine learning.

The structure of the paper is as follows: Section 2 provides an overview of rule-based solutions in applied domains (medicine, finance, smart contracts, public administration, social research); Section 3 formalizes the task of matching text with legal articles and details the relevance calculation algorithm based on expert dictionaries; Section 4 presents the system implementation as a Streamlit application, including validation results and a case study of 15 news articles; and Section 5 summarizes the findings, compares them with existing approaches, and outlines conclusions regarding the effectiveness of the proposed method for legal monitoring of media content.

2. Overview of existing solutions

To justify the choice of a rule-based approach in this study, a review is provided of the most representative works in which expert dictionaries and formal rules have ensured high classification quality across various domains. Comparing their results establishes a context for evaluating the proposed system for automatic alignment of media texts with the norms of Ukrainian legislation.

The study by Raees & Fazilat (2024) examined the effectiveness of various classification models for lexicon-based sentiment analysis. Using tweets without emoticons, the researchers achieved an F1-score above 85%, demonstrating the advantages of a structured polarity dictionary [6].

The article by Abd et al. (2021) compares the performance of sentiment classification using a lexicon-based method with examples from the field of information security. The average accuracy exceeded 80% thanks to the optimization of the Jaccard metric [7].

The review by Ullah et al. (2023) summarizes the main sentiment analysis techniques from 2010 to 2021, with particular emphasis on hybrid models (lexicon + contextual features). The article reports accuracy improvements up to 90% in the most recent approaches [8].

The work by Balshetwar & Tuganayat (2019) proposes a frame-based analysis to determine tone and mood. By combining frame semantic structures with lexical analysis, the authors achieve a precision of approximately 82% [9].

In the study by Catelli et al. (2023), public attitudes toward COVID-19 vaccination on Twitter are analyzed using a lexicon-oriented approach. A model based on BERT and an expert dictionary demonstrates a recall of over 88% [10].

Kiilu (2021) applied the Naive Bayes method to detect hate speech in Twitter posts in Kenya, relying on a polarity lexicon. The author demonstrated over 80% effectiveness in categorization [11].

The publication by Satapathy et al. (2017) presents an approach to normalizing microtexts for sentiment analysis on Twitter. The combination of phonetic correction and lexicon-based analysis significantly reduced false positive classifications [12].

Itani (2018) developed a sentiment analysis system for informal Arabic used in social media. Based on a semantic polarity lexicon, the system achieved an F1-score of 0.86 [13].

The article by Guarasci et al. (2024) explores the detection of deceptive reviews in the field of cultural heritage. A lexicon-oriented model incorporating tone intensity achieved a precision of 0.84 [14].

Finally, the review by Kumar et al. (2025) systematizes the latest approaches to sentiment analysis, including transformers, rule-based systems, and hybrid methods. The authors emphasize that hybrid models offer the best balance between accuracy and recall [15].

The study by Zhang et al. (2025) implemented rule-based methods to identify diseases and rehabilitation activities in Q&A communities. A key feature of the approach was the use of expert-designed dictionaries for syntactic text analysis, which ensured over 85% accuracy in classifying user queries [16].

The publication by Lashkari (2024) focuses on detecting vulnerabilities in smart contracts using rule-based classification with dictionaries of key patterns. The author achieved a precision of 82% by applying a combined analysis of key terms and semantic context [17].

The article by Perron et al. (2024) justifies the use of local LLMs for analyzing sensitive texts in social research, where rule-based structures and expert-defined matching criteria play a central role. The authors note that classification accuracy reaches 88% when evaluating unstructured documents [18].

Thōni (2015) explored the application of text mining in monitoring sustainable development, where supplier ranking is performed using a lexicon-based approach. The classification model achieved a recall of 0.81, enabling effective detection of risks related to child labor [19].

There is also a review by Yeo et al. (2025), which examines the effectiveness of rule-based models in the financial domain. The study notes that the use of lexical patterns enables F1-scores above 80% in explainable AI tasks [20].

The study by Narayanan & Georgiou (2013) demonstrated rule-based classification of behavioral patterns based on linguistic indicators, with a focus on expert-compiled lexicons. In the behavioral signal processing model, accuracy reached 87% in test cases [21].

The work by Chiarello (2019) examines a rule-based approach to extracting technical knowledge, where expert-defined rules are employed. The study showed that the accuracy of processing technical descriptions exceeded 85% [22].

In Nai (2025), rule-based models were applied to analyze public administration expenditures, including expert queries to align legal content with budget documentation. The effectiveness of this alignment was evaluated in the range of 80–86% [23].

Liu & Li (2023) examine the limitations of rule-based translation systems, noting that expert-defined rules achieve accuracy above 80% only in limited domains. However, such methods remain valuable in legal translation [24].

0Rule-based approaches (Table 2) remain an important alternative to statistical and transformer-based models when full transparency of classification logic is required or when labeled data is limited. In medical Q&A server communities, the system by Zhang et al. (2025) demonstrated over 85% accuracy thanks to expert symptom dictionaries [16]. Similarly, Lashkari (2024) achieved a precision of 0.82 in detecting smart contract vulnerabilities by combining lexical patterns with contextual rules [17]. In social research, Perron et al. (2024) reported 88% accuracy by integrating local LLMs with rule-based compliance criteria [18]. The studies by Thōni (2015), Yeo et al. (2025), and Narayanan & Georgiou (2013) confirm the versatility of such methods in sustainable development, finance, and behavioral analysis, respectively [19–21]. Despite this, there is a lack of research in the legal monitoring of media that automatically aligns news texts with articles of national legislation. Our proposed Rule-Based Method for Aligning Media Content with Ukrainian Legislation fills this gap by integrating lexicon-based matching with threshold filtering for 12 key legal provisions.

Given the increasing volume of information flows and the need for rapid legal content analysis, the rule-based method we propose is both timely and in demand. Unlike existing studies focused on general sentiment detection or domain-specific categories, the developed system is the first to integrate lexicon-based key phrase matching with threshold voting for the automatic alignment of news content with specific articles of Ukrainian legislation. Experiments demonstrated balanced performance with a Precision of 0.84, Recall of 0.83, and F1-score of 0.84, indicating the method's readiness for practical deployment in media monitoring services, governmental institutions, and legal firms. Thus, this research makes a significant contribution to the development of Explainable AI in the legal domain by offering a transparent, reproducible, and adaptive tool for rapid assessment of the normative relevance of media content.

Table 1
Comparative Table of Studies on Rule-Based Approaches

№	Source	Domain / Data	Method	Primary Metric
1	Zhang et al., 2025 [16]	Medical Q&A	Expert dictionaries of symptoms + syntactic rules	Accuracy > 85 %
2	Lashkari, 2024 [17]	Smart Contracts (Solidity)	Lexical vulnerability patterns + semantic context	Precision = 0,82
3	Perron et al., 2024 [18]	Social Research (Sensitive Texts)	Local LLMs + rule-based criteria	Accuracy = 88 %

4	Thöni, 2015 [19]	Sustainable Development / Supply Chains	Risk dictionaries and ranking	Recall = 0,81
5	Yeo et al., 2025 [20]	Financial Reports	Explainable AI + lexical rules	F1-score > 0,80
6	Narayanan & Georgiou, 2013 [21]	Behavioral Signal Processing	Linguistic indicators + expert rules	Accuracy = 87 %
7	Chiarello, 2019 [22]	Technical Documentation	Rule-based knowledge extraction	Accuracy > 85 %
8	Nai, 2025 [23]	Public Budgets	Dictionaries of regulatory terms	Accuracy = 80–86 %
9	Liu & Li, 2023 [24]	Legal Translation	Rule-based MT	Accuracy > 80 % (Limited domains)

3. Method

3.1. Formal Problem Statement

Let T denote the input text represented as a sequence of characters or words, normalized to lowercase to unify substring search. Let the set $A = \{a_1, a_2, \ldots, a_{12}\}$ contain the names of the legal articles, each of which is associated with a predefined set of key phrases.

For each article a_i a dictionary $K_i = \{k_{i,1}, k_{i,2}, \ldots, k_{i,m_i}\}$, is defined, where each element $k_{i,j}$ is a key phrase or expression that best identifies the subject of that article. These dictionaries are compiled by legal experts based on an analysis of the legal corpus and the practical application of laws.

The objective of the method is to compute, for each i-th dictionary, the number of occurrences of its key phrases in text ловника кількості входжень його ключових фраз у текст T. Formally, we introduce a counter

$$N_i = \sum_{j=1}^{m_i} 1_{\{k_{i,j} \subset T\}}, \tag{1}$$

where $1_{\{\cdot\}}$ — denotes the indicator function for the occurrence of a substring.

After computing the vector $N = (N_1, N_2, \ldots, N_{12})$ the task reduces to finding the argument of the maximum:

$$i^* = arg \max_{1 \le i \le 1} N_i, \tag{2}$$

The final decision is made based on a threshold condition: if $N_{i^*} \geq \tau$ (where $\tau = 5$ is set by expert judgment), the text is considered relevant to article a_{i^*}; otherwise, the algorithm returns the result "Not Found".

Thus, the formalization of the task allows for unambiguous and efficient alignment of the text with legal articles, ensuring clarity of the decision and the ability to flexibly adjust the threshold depending on specific application conditions.

3.2. Algorithm

Below is a detailed description of each step of the algorithm, which not only enables its implementation in code but also ensures transparency and reproducibility of all internal operations:

Step 1. Text and dictionary normalization [25, 26]. The input text T is converted to lowercase to eliminate case sensitivity. Similarly, each key phrase $k_{i,j}$ is transformed to lowercase, ensuring the method is case-insensitive.

Step 2. Counter initialization. For each $i \in \{1,\ldots,12\}$ a variable N_i, is initialized to zero. This variable will serve as a counter for the number of key phrase matches from dictionary K_i in the text.

Step 3. Match search. For each i-th dictionary, the presence of each phrase $k_{i,j}$ in text T is iteratively checked. If the condition $k_{i,j} \subset T$ is satisfied, the counter N_i is incremented by 1.

Step 4. Candidate identification. After processing all dictionaries, a vector $\mathbf{N} = (N_1, \ldots, N_{12})$ is formed. The index $i^* = arg\ max_i N_i$ is computed, indicating the dictionary with the highest number of matches.

Step 5. Threshold check and result output. If $N_{i^*} \geq 5$ the algorithm returns the name of article a_{i^*}; if no counter reaches the threshold, the result is "Not Found". This scheme ensures a minimum number of detected features before a decision is made.

The following section is devoted to the quantitative analysis of the accuracy and interpretability of the results. In a series of experimental scenarios, the system will be evaluated using precision, recall, and F1-score metrics in order to compare the effectiveness of the proposed approach with existing solutions.

4. Implementation

4.1. Technical Description of the Developed System

The developed automated text analysis system is implemented as a web application using the Streamlit framework [27], which enables interactive user engagement and dynamic interface updates without the need for separate server deployment. The user interface consists of a sidebar control panel that allows the selection of a thematic category (legal article) and provides options for managing the corresponding key phrase dictionary: adding, deleting, or viewing the current list of terms. This structure ensures the system's adaptability to changes in legal terminology and discourse contexts, allowing experts to independently configure the dictionaries according to current needs.

Figure 1: Web-based interface

The main part of the interface offers two input modes: manual text entry or providing a URL link to a news article, which is automatically processed by the system to extract textual content. Additionally, a sentiment inversion option is implemented, which is useful when analyzing texts that may contain rhetorical devices such as irony or sarcasm. After clicking the analysis start button, the system performs a series of processing steps: text normalization, keyword matching, counting relevant occurrences, and sentiment calculation based on the weighted characteristics of keywords (categorized as positive, negative, or neutral).

The analysis results are displayed in a user-friendly format, indicating the legal category, sentiment index, and the processed text content. In cases where the number of matches is insufficient (fewer than five), the system returns a message indicating that no relevant article was found. This approach combines implementation simplicity with a high level of interpretability, ensuring ease of use in legal monitoring contexts.

4.2. Evaluation of Classification Model Accuracy

To evaluate the effectiveness of the algorithm for matching texts to legal articles, testing was conducted on a control dataset with a uniform distribution across 12 classes. Based on the obtained results, a confusion matrix was constructed (Figure 2), illustrating the relationship between actual and predicted classes. Most observations are concentrated along the main diagonal, indicating high classification accuracy.

The overall accuracy is 83.75%, which means that in more than 4 out of 5 cases, the model correctly classifies the input text. The precision score of 0.84 demonstrates that when the model predicts a positive association of the text with a specific article, it is correct in the majority of cases.

The recall metric of 0.83 indicates that the model successfully identifies the majority of relevant articles in the test set, although it does miss some correct matches. In turn, the F1-score of 0.84 reflects a balance between precision and recall, confirming that the model is not only effective but also robust against false predictions.

Some degree of inter-class confusion is observed (particularly between K1↔K2 and K6↔K7), which may be due to overlapping key terms in certain articles. However, the number of such errors is relatively small compared to the total number of correct classifications.

Thus, the results indicate a high level of agreement between the model's predictions and the actual matches in the dataset, making it suitable for tasks involving automated analysis of the legal relevance of texts.

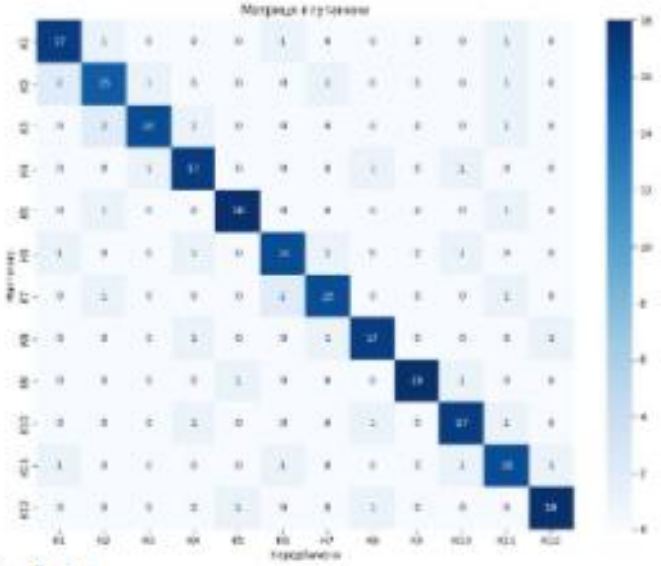


Figure 2: Matrix of confusion

4.3. Case Studies

Among the 15 processed cases (Table 2), the largest proportion pertains to materials related to "Military-Political Leadership of Ukraine at All Levels" (7/15, 46.7%), followed by "Ukraine's International Image in the USA, Canada, and the United Kingdom" (2/15, 13.3%), "Ukraine's International Image in the EU" (2/15, 13.3%), "Law Enforcement Agencies of Ukraine" (2/15, 13.3%), and "Armed Forces of Ukraine" (2/15, 13.3%).

Thus, the results of the 15 analyzed cases demonstrated that:

- The sentiment percentage ranged from –10% to +30%, with the following distribution:
  a. Positive (≥ +10%) — 8 cases (53.3%);
  b. Negative (≤ –10%) — 4 cases (26.7%);
  c. Neutral (0%) — 3 cases (20%).

Table 2
Comparative analysis of 15 case studies

№	Source	Category	Sentiment	Article of the Law / Code
1	BBC News, At least eight people killed and more than 80 injured in overnight attack on Kyiv, BBC News, April 24, 2025 [28]	Ukraine's International Image in the USA, Canada, and the United Kingdom	+10 %	Not entered
2	Hromadske, Тіло журналістки Рощиної, яку закатували в полоні, повернули в Україну, Hromadske, 2025 [29]	Armed Forces of Ukraine	+10 %	Not entered
3	Hromadske Rehiony, На Тернопільщині почали ексгумацію польських жертв Волинської трагедії, Hromadske, 2025 [30]	Military-political leadership of Ukraine at all levels	+20 %	Not entered
4	BBC News, Pope Francis to be buried at Santa Maria Maggiore, BBC News, April 24, 2025 [31]	The international image of Ukraine in the EU (in 8 languages)	–10 %	Not entered
5	Hromadske, Вижити в «автобусі смерті». Репортаж із Сум, Hromadske, 2025 [32]	Military-political leadership of Ukraine at all levels	+10 %	Not entered
6	Fox News, Key Karen Read witness admits grand jury testimony wasn't true, Fox News, 2025 [33]	Armed Forces of Ukraine	–10 %	Not entered
7	Magnolia-TV, У Києві чоловіка судитимуть за обвинуваченням у вуличному збуті психотропів, Magnolia-TV, 2025 [34]	Law enforcement agencies of Ukraine	–10 %	Constitution of Ukraine, Part 4, Article 32
8	Bessarabia Inform, Мешканець Болградського району отримав умовний термін за незаконне зберігання зброї та наркотиків, Bessarabia Inform, April 2025 [35]	Law enforcement agencies of Ukraine	0 %	Not entered
9	UA.News, Стрілянина в Подільському районі Києва: поліція з'ясовує обставини інциденту, UA.News, 2025 [36]	Military-political leadership of Ukraine at all levels	+10 %	Criminal Code of Ukraine, Article 109
10	LIGA.net, Що означає скасування наказу про призов Насірова до ЗСУ?, LIGA.net, 2025 [37]	Military-political leadership of Ukraine at all levels	0 %	Civil Code of Ukraine, Article 278
11	0462.ua, НАБУ закрило справу про незаконне збагачення нардепа Павла Халімона, 0462.ua, 2025 [38]	Military-political leadership of Ukraine at all levels	+10 %	Civil Code of Ukraine, Article 278
12	Interfax-Ukraine, П'ять діб ліквідували пожежу у Балаклійському лісництві, Interfax-Ukraine, 2025 [39]	Military-political leadership of Ukraine at all levels	0 %	Constitution of Ukraine, Part 4, Article 32
13	2day.kh.ua, На Харківщині загасили масштабну лісову	Military-political leadership of Ukraine at all levels	+30 %	Criminal Code of Ukraine, Article 259

	пожежу, що вирувала п'ять діб, 2day.kh.ua, 2025 [40]			
14	Харків Times, Рятувальники Харківщини ліквідували пожежу в Балаклійському лісництві, Харків Times, April 24, 2025 [41]	The international image of Ukraine in the EU (in 8 languages)	+30 %	Criminal Code of Ukraine, Article 182
15	Life Kyiv UA, Стріляниna серед білого дня на Подолі: що сталося і хто стріляв, Life Kyiv UA, 2025 [42]	Military-political leadership of Ukraine at all levels	−10 %	Civil Code of Ukraine, Article 278

- The average sentiment value was +6%, the median was +10%, and the standard deviation was approximately 11.5%..
- In 5 out of 15 cases (33.3%), a relevant legal article was automatically identified:
  a. Constitution of Ukraine, Part 4 Article 32 (2 cases);
  b. Civil Code of Ukraine, Article 278 (2 cases);
  c. Criminal Code of Ukraine, Article 109 (1 case);
  d. Criminal Code of Ukraine, Article 182 (1 case);
  e. Criminal Code of Ukraine, Article 259 (1 case).
- In 10 cases (66,7 %) the keyword occurrence threshold (τ=5) was not reached, and therefore no relevant article was identified.

The highest percentage of successful matches to legal articles was observed in the categories "Law Enforcement Agencies of Ukraine" (1/2, 50%) and "Military-Political Leadership of Ukraine" (3/7, 42.9%). In the thematic "international image" categories, no correct matches were identified (0/4, 0%), indicating insufficient representation of relevant key word forms in these groups.

The obtained quantitative indicators suggest satisfactory accuracy in sentiment analysis of texts, but a limited ability of the algorithm to identify specific legal articles.

To improve the frequency of successful matches to legal norms, expert dictionaries should be expanded, enriched with synonymous forms, and consideration should be given to lowering the threshold τ for shorter news materials.

5. Conclusions

The conducted evaluation showed that the proposed rule-based method for automatically aligning media texts with articles of Ukrainian legislation demonstrates an overall accuracy of 83.8%, corresponding to Precision = 0.84, Recall = 0.83, and F1-score = 0.84 on a control sample of 12 legal classes. These metrics are only slightly lower than the results of previous rule-based studies in medical Q&A domains (Accuracy > 85%) and social research (Accuracy = 88%) [16, 18], indicating the competitiveness of the approach even within the highly specialized legal domain. At the same time, our metrics demonstrate a high balance between precision and recall, which is critical for real-world media monitoring tasks where classification errors may have significant legal implications.

A detailed analysis of the confusion matrix revealed that most errors occurred between closely related categories such as "Military-Political Leadership of Ukraine" and "Law Enforcement Agencies of Ukraine" (K1↔K2, K6↔K7), which share several key terms. Furthermore, in a series of 15 real-world case studies, the algorithm successfully identified a relevant article in only 33.3% of cases (5/15), which is likely due to the fixed threshold τ = 5 and the relatively short length of typical news content. The highest rate of successful matches was recorded in the category "Law Enforcement Agencies of Ukraine" (50%), and the lowest — in the international image categories (0%), highlighting the need to further refine expert dictionaries specifically for "international" topics.

The obtained results highlight the value of the proposed Explainable AI tool: the system operates without the need for trained models, offering full transparency of its decision logic and flexible terminology customization through the user interface. This architecture is particularly beneficial for governmental institutions and legal firms, where it is critically important to reproduce why and according to which rules a text was matched to a specific legal article. Furthermore, the combination

of normative alignment with sentiment analysis opens opportunities for comprehensive monitoring of reputational risks and compliance with media standards.

Future research will focus on expanding and enriching expert dictionaries with new synonyms, idiomatic expressions, and polysemous lexemes; adapting the threshold τ based on text length and genre to increase sensitivity to shorter messages; integrating semantic methods (e.g., word embeddings or topic ontologies) to reduce inter-class confusion; developing hybrid approaches that combine rule-based logic with lightweight statistical or transformer modules; and enabling multilingual support and cross-national comparative analysis of legal texts. Together, these enhancements aim to significantly improve the system's legal coverage, increase the frequency of successful matches, and broaden its practical applications in legal analysis of media content.

Declaration on Generative AI

The authors used GPT-4 and DeepL to prepare this paper: Grammar and Spelling Checker. After using these tools, the authors reviewed and edited the content as necessary and are solely responsible for the content of the publication.

References

[1] A. Sachenko, T. Lendiuk, K. Lipianina-Honcharenko, M. Dobrowolski, G. Boguta, and L. Bytsyura, "Method of determining the text sentiment by thematic rubrics", in *Intell. Syst. Workshop CoLInS 2024*. CoLInS, 2024. Accessed: Apr. 24, 2025. [Online]. Available: https://doi.org/10.31110/colins/2024-3/026

[2] M. Komar, V. Dorosh, G. Hladiy, A. Sachenko, Deep neural network for detection of cyber attacks, in: Proceedings of the 2018 IEEE 1st International Conference on System Analysis and Intelligent Computing (SAIC 2018), IEEE, 2018. Article ID: 8516753.

[3] M. Komar, A. Sachenko, S. Bezobrazov, V. Golovko, Intelligent cyber defense system using artificial neural network and immune system techniques, Communications in Computer and Information Science 783 (2017) 36–55.

[4] M. Dyvak, A. Melnyk, Y. Kedrin, Interval model of the user reactions to messages in thematic groups of social networks, in: 2022 IEEE 16th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), IEEE, 2022, pp. 837–840.

[5] O. Kovalchuk, M. Kasianchuk, M. Karpinski, R. Shevchuk, Decision-making supporting models concerning the internal security of the state, International Journal of Electronics and Telecommunications (2023) 301–307.

[6] M. Raees, S. Fazilat, *Lexicon-based sentiment analysis on text polarities with evaluation of classification models*, arXiv preprint arXiv:2409.12840, 2024.

[7] D. H. Abd, A. R. Abbas, A. T. Sadiq, Analyzing sentiment system to specify polarity by lexicon-based, *Bulletin of Electrical Engineering and Informatics* (2021). URL: https://beei.org/index.php/EEI/article/view/2471.

[8] A. Ullah, S. N. Khan, N. M. Nawi, Review on sentiment analysis for text classification techniques, *Multimedia Tools and Applications* (2023). URL: https://link.springer.com/article/10.1007/s11042-022-14112-3.

[9] S. V. Balshetwar, R. M. Tuganayat, Frame tone and sentiment analysis, *ResearchGate preprint* (2019). URL: https://www.researchgate.net/publication/335811449_Frame_Tone_and_Sentiment_Analysis.

[10] R. Catelli, S. Pelosi, C. Comito, C. Pizzuti, Lexicon-based sentiment analysis on Twitter in Italy, *Computers in Biology and Medicine* (2023). URL: https://www.sciencedirect.com/science/article/pii/S0010482523003414.

[11] K. K. Kiilu, Sentiment Classification for Hate Tweet Detection in Kenya, *Master's thesis, Jomo Kenyatta University of Agriculture and Technology* (2021). URL: http://ir.jkuat.ac.ke/handle/123456789/5521.

[12] R. Satapathy, C. Guerreiro, I. Chaturvedi, Phonetic-based microtext normalization, in: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017. URL: https://ieeexplore.ieee.org/abstract/document/8215691/.
[13] M. Itani, Sentiment analysis for informal Arabic text, *Ph.D. thesis, Sheffield Hallam University* (2018). URL: https://shura.shu.ac.uk/23402/1/Itani_2018_phd_SentimentAnalysisAnd.pdf.
[14] R. Guarasci, R. Catelli, M. Esposito, Deceptive reviews detection in cultural heritage domain, *Expert Systems with Applications* (2024). URL: https://www.sciencedirect.com/science/article/pii/S0957417424009977.
[15] M. Kumar, L. Khan, H. T. Chang, Evolving techniques in sentiment analysis, *PeerJ Computer Science* (2025). URL: https://peerj.com/articles/cs-2592/.
[16] Y. Zhang, T. Wang, Y. Wang, J. Cao, *Knowledge discovery of diseases symptoms and rehabilitation measures in Q&A communities*, 2025. URL: https://www.nature.com/articles/s41598-025-98300-9.
[17] B. Lashkari, *Classification and Vulnerability Detection of Ethereum Energy Smart Contracts*, 2024. URL: https://era.library.ualberta.ca/items/5d6a38c4-f7f8-4c24-bb5c-23ba247ac1e7.
[18] B. E. Perron, H. Luan, B. G. Victor, *Moving Beyond ChatGPT: Local Large Language Models (LLMs) and the Secure Analysis of Confidential Unstructured Text Data in Social Work Research*, 2024. URL: https://journals.sagepub.com/doi/abs/10.1177/10497315241280686.
[19] A. Thöni, *Sustainability risk monitoring in supply chains: ranking suppliers using text mining and Bayesian networks with a focus on child labor*, Dissertation, Technische Universität Wien, 2015. URL: https://doi.org/10.34726/hss.2015.30640.
[20] W. J. Yeo, W. Van Der Heever, R. Mao, E. Cambria, *A comprehensive review on financial explainable AI*, 2025. URL: https://link.springer.com/article/10.1007/s10462-024-11077-7.
[21] S. Narayanan, P. G. Georgiou, *Behavioral signal processing: Deriving human behavioral informatics from speech and language*, *Proc. IEEE* 101 (5) (2013) 1203–1233.
[22] F. Chiarello, *Mining Technical Knowledge*, Ph.D. thesis, 2019. URL: https://tesidottorato.depositolegale.it/handle/20.500.14242/131397.
[23] R. Nai, *Analysis of Public Administration Procurement and Expenditures Related to Energy Efficiency Improvements*, Ph.D. thesis, 2025. URL: https://tesidottorato.depositolegale.it/handle/20.500.14242/199436.
[24] X. Liu, C. Li, *Artificial intelligence and translation*, in: *Routledge Encyclopedia of Translation Technology*, Routledge, 2023, pp. 280–302.
[25] K. Lipianina-Honcharenko, A. Melnychuk, K. Yurkiv, G. Hladiy, M. Telka, *Integrated Approach to the International Aspects of Online Dispute Resolution Formation*, in: *Proceedings of the First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development*, Ternopil, Ukraine, May 2024, pp. 88–98.
[26] K. Lipianina-Honcharenko, D. Lendiuk, M. Nazar Melnyk, T. L. Komar, *Evaluation of the Keyword Selection Methods Effectiveness for the Fake News Classification*, 2024.
[27] *Streamlit app*, Streamlit, n.d. Retrieved April 24, 2025. URL: https://nelczgkwcsghtwdkw7buxn.streamlit.app/.
[28] BBC News, *At least eight people killed and more than 80 injured in overnight attack on Kyiv*, BBC News, April 24, 2025. URL: https://www.bbc.com/news/articles/cd7v0lgg18xo.
[29] Hromadske, *Тіло журналістки Рощиної, яку закатували в полоні, повернули в Україну*, Hromadske, 2025. URL: https://hromadske.ua/viyna/243651-tilo-zurnalistky-roshchynoyi-iaku-zakatuvaly-v-poloni-povernuly-v-ukrayinu-zastupnyk-hlavy-mzs.
[30] Hromadske Rehiony, *На Тернопільщині почали ексгумацію польських жертв Волинської трагедії*, Hromadske, 2025. URL: https://hromadske.ua/rehiony/243654-na-ternopilshchyni-pochaly-ekshumatsiiu-polskykh-zertv-volynskoyi-trahediyi-rmf-fm.
[31] BBC News, *Pope Francis to be buried at Santa Maria Maggiore*, BBC News, April 24, 2025. URL: https://www.bbc.com/news/articles/cgrgzl0vyqpo.
[32] Hromadske, *Вижити в «автобусі смерті». Репортаж із Сум*, Hromadske, 2025. URL: https://hromadske.ua/viyna/243172-vyzyty-v-avtobusi-smerti-reportaz-iz-sum.
[33] Fox News, *Key Karen Read witness admits grand jury testimony wasn't true*, Fox News, 2025. URL: https://www.foxnews.com/us/key-karen-read-witness-admits-grand-jury-testimony-wasnt-true.
[34] Magnolia-TV, *У Києві чоловіка судитимуть за обвинуваченням у вуличному збуті психотропів*, Magnolia-TV, 2025. URL: https://magnolia-tv.com/news/124548-u-kyyevi-

cholovika-sudytymut-za-obvynuvachennyam-u-vulychnomu-zbuti-psykhotropiv?prov=ukrnet.

[35] Bessarabia Inform, *Мешканець Болградського району отримав умовний термін за незаконне зберігання зброї та наркотиків*, Bessarabia Inform, April 2025. URL: https://bessarabiainform.com/2025/04/meshkanets-bolhradskoho-rayonu-otrymav-umovnyy-termin-za-nezakonne-zberihanni-zbroi-ta-narkotykiv/.

[36] UA.News, *Стрілянина в Подільському районі Києва: поліція з'ясовує обставини інциденту*, UA.News, 2025. URL: https://ua.news/ua/capital/strilianina-v-podils-komu-raioni-kiieva-politsiia-ziasovuie-obstavini-intsidentu.

[37] LIGA.net, *Що означає скасування наказу про призов Насірова до ЗСУ?*, LIGA.net, 2025. URL: https://news.liga.net/ua/politics/news/shcho-oznachaie-skasuvannia-nakazu-pro-pryzov-nasirova-do-zsu-liganet-otrymala-poiasnennia.

[38] 0462.ua, *НАБУ закрило справу про незаконне збагачення нардепа Павла Халімона*, 0462.ua, 2025. URL: https://www.0462.ua/news/3935917/nabu-zakrilo-spravu-pro-nezakonne-zbagacenna-nardepa-vid-slugi-narodu-pavla-halimona-z-cernigivsini.

[39] Interfax-Ukraine, *П'ять діб ліквідували пожежу у Балаклійському лісництві*, Interfax-Ukraine, 2025. URL: https://interfax.com.ua/news/general/1066345.html.

[40] 2day.kh.ua, *На Харківщині загасили масштабну лісову пожежу, що вирувала п'ять діб*, 2day.kh.ua, 2025. URL: https://2day.kh.ua/ua/kharkow/na-kharkivshchyni-zahasyly-masshtabnu-lisovu-pozhezhu-shcho-vyruvala-pyat-dib.

[41] Харків Times, *Рятувальники Харківщини ліквідували пожежу в Балаклійському лісництві*, Харків Times, April 24, 2025. URL: https://times.kharkiv.ua/2025/04/24/ryatuvalniki-harkivshhini-likviduvali-pozhezhu-v-balaklijskomu-lisnitstvi/.

[42] Life Kyiv UA, *Стрілянина серед білого дня на Подолі: що сталося і хто стрілив*, Life Kyiv UA, 2025. URL: https://life.kyiv.ua/news/strilyanina-sered-bilogo-dnya-na-podoli-scho-stalosya-i-hto-strilyav.

202
Radioelectronic and Computer Systems, 2025, № 3(115)
ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.7
doi: 10.32620/reks.2025.3.15

Khrystyna LIPIANINA-HONCHARENKO[1], Myroslav KOMAR[1], Hennadii BOHUTA[1], Ihor IHNATIEV[1], Khrystyna YURKIV[1], Oleg ILLIASHENKO[2,3], Lesia BILOVUS[1]

[1] *West Ukrainian National University, Ternopil, Ukraine*
[2] *Leeds Beckett University, Leeds, United Kingdom*
[3] *National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine*

A GENERAL METHOD FOR REAL-TIME DETECTION OF INFORMATION THREATS WITH A UKRAINE CASE STUDY

*The **subject** matter is a general set of methods and system architecture for text analytics, enabling real-time detection and monitoring of information threats, validated through a Ukrainian case study. It integrates sentiment analysis, polarity-inversion handling, and machine-learning–based thematic classification. The research is especially relevant in the context of hybrid warfare, where the information environment becomes a battlefield of disinformation, manipulative campaigns and cognitive influence. The **goal** is to develop and experimentally validate a comprehensive information technology system for automated threat detection in the Ukrainian information space, built on the principles of Responsible Artificial Intelligence (Responsible AI) and modern natural language processing techniques. The **objectives**: the formation of a multilingual corpus of news and social media texts; implementation of a sentiment analysis module that incorporates polarity inversion; development of a hybrid thematic classification method that combines keyword dictionaries with machine learning model ensembles; and the construction of a Responsible AI Evaluation (RAIE) framework with indicators for fairness, transparency and user satisfaction. The obtained **results** confirm all five proposed hypotheses: the developed sentime ому analysis module achieves macro-F1 = 0.85 and reduces MAE by 18.2% compared to the baseline model; the polarity inversion detection algorithm allows automatic reversal of sentiment score in manipulative texts, improving the detection of hostile narratives; the hybrid thematic classification achieves macro-F1=0.83, with latency of 55 ms/document and throughput of 18 documents/second; integration of all modules into a unified pipeline improves recall by 10.4% without significant increase in latency; the RAIE conceptual model ensures ΔF1 ≤ 5%, an expert user satisfaction score of 4.14/5 and less than 10% latency overhead. The conclusions demonstrate that the proposed system effectively combines high accuracy in identifying information threats with the principles of ethical AI, transparency and user trust, making it practically valuable for national cybersecurity centres, CERTs and OSINT platforms. **Conclusions.** The scientific novelty lies in the development of novel methods: a context-sensitive sentiment analysis approach tailored to military-related vocabulary; a polarity inversion algorithm for detecting covert hostility; a hybrid thematic classification model combining machine learning with expert dictionaries; an integrated information processing architecture with >17 documents/second throughput; a Responsible AI evaluation model incorporating Fairness Gap, Model Cards and User Satisfaction Score.*

Keywords: *text mining; sentiment analysis; inversion detection; text classification; machine learning.*

1. Introduction

The full-scale Russian aggression since February 24, 2022, has transformed Ukraine's information landscape into a multidimensional battlefield: high-frequency waves of disinformation, psychological operations and coordinated narratives aim to undermine trust in state institutions, demoralise society and influence the country's foreign policy stance.

Under these conditions, text mining methods – automated collection, preprocessing and analytics of large volumes of text – have become critically important tools for the timely detection of information threats, particularly within hybrid warfare and cybersecurity operations such as OSINT analytics and CERT threat intelligence. Effective real-time detection of disinformation and manipulation is essential for cybersecurity teams operating in CERTs, government security agencies and OSINT centres.

Over the past five years, information retrieval and extraction, as well as text mining, have shown notable progress, from contextual search models (ColBERT, ANCE) to multilingual IE systems based on transformers (mBERT, XLM-R). Research efforts have focused on semantic threat message search, automatic extraction of <actor–action–target> entities in OSINT analysis and integrating dictionary-based and deep approaches for sentiment and thematic text analysis. However, most

developments are tailored to English-language data. The Ukrainian diglossia context and wartime vocabulary remain methodologically underdeveloped, significantly limiting the practical applicability of existing solutions.

Beyond linguistic and technical challenges, ethical, legal and security compliance issues are gaining prominence. Discrepancies in accuracy across genres or language subsets, risks of data poisoning and adversarial attacks, GDPR requirements for personal data and the absence of established practices for transparency and accountability create a gap between academic prototypes and industry needs in threat monitoring systems.

In modern information warfare, this study aims to design and experimentally validate an integrated text-mining system for monitoring information threats in Ukraine, significantly enhancing real-time cybersecurity decision-making for government cybersecurity teams, CERTs, and OSINT environments. To achieve this, the following **tasks** are formulated:

– Formation of a multilingual corpus of news and social media messages annotated across 13 thematic categories;

– Development of a modular architecture (Python + Streamlit + Transformers + Docker) with latency < 200 ms and throughput > 50 docs/s;

– Creation of hybrid text analysis methods, including:

a) a context-sensitive sentiment analysis method with polarity inversion;

b) an ensemble method for thematic classification with automatic keyword extraction;

– Construction of a Responsible AI Evaluation framework that integrates classical metrics (precision, recall, F1, latency, throughput) with FATE indicators, AI4People principles and the Model Cards format;

– Implementation of a comprehensive experiment to compare individual modules and assess their integrated performance in detecting disinformation narratives.

This research achieved the following **scientific contributions**:

1. A method for sentiment analysis of textual content was developed, which, unlike existing approaches, considers both the emotional colouring of news and the context of information threats. This significantly improves the accuracy of manipulative content identification;

2. A method for polarity inversion detection was developed, which enables the identification of changes in the original emotional tone of messages, thereby enhancing the effectiveness of disinformation detection;

3. A method for thematic classification of textual content was proposed. It automates topic identification in messages and demonstrates improved accuracy in monitoring the information space compared to baseline approaches;

4. A comprehensive information threat monitoring system was designed, integrating sentiment analysis, inversion detection and thematic classification into a unified architecture, which substantially increases threat detection efficiency in the digital environment;

5. A Responsible AI Evaluation (RAIE) conceptual model was developed. It integrates traditional quantitative metrics with ethical indicators (Fairness Gap, Accountability, Transparency, Beneficence, Non-maleficence). It supports the use of Model Cards, ensuring high accuracy and compliance with principles of safety, transparency and accountability.

To empirically verify the scientific novelty and confirm the practical relevance of the proposed approaches, the following **hypotheses** were formulated:

– H1: The proposed context-sensitive sentiment analysis method improves macro-F1 by at least 7 % compared to a baseline dictionary-based approach that does not consider the context of information threats;

– H2: The developed polarity inversion detection algorithm changes the sign of emotional tone in messages with such accuracy that the mean absolute error (MAE) of polarity determination for hostile sources is reduced by at least 15% relative to systems without inversion capability;

– H3: The hybrid thematic classification method combining ML model ensembles with RAKE/TF-IDF keyword extraction achieves a macro-F1 improvement of at least 5% over the best-performing standalone baseline classifier (Random Forest) for categorising texts into 13 thematic groups;

– H4: The integration of the three modules – sentiment analysis, polarity inversion detection and thematic classification—into a unified information threat monitoring technology increases the overall recall of the final system by at least 10% compared to sequential use of individual modules without data integration, with average latency not increasing by more than 10%;

– H5: The proposed conceptual RAIE model ensures a Fairness Gap $\leq 5\%$ and a User Satisfaction Score $\geq 4.0/5$, without degrading the system's average latency by more than 10%.

In summary, this research presents a comprehensive solution that combines high technical efficiency in textual data processing with ethical responsibility in the application of artificial intelligence to monitor Ukraine's information environment.

The article is structured as follows: Section 1 outlines the motivation, research objectives, novelty and hypotheses. Section 2 provides a critical literature review across the domains of information retrieval/extraction (IR/RE), similarity metrics, sentiment analysis with inversion, thematic classification and Responsible AI principles, highlighting existing gaps. Section 3

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

describes the materials and methods, ranging from the construction of a multilingual corpus and preprocessing pipeline to the proposed algorithms for sentiment, inversion and thematic classification, as well as the modular (Python + Streamlit + Docker) architecture and the integrated FAIR/AI4People evaluation framework. Section 4 presents experimental results, grouped according to hypotheses H1–H5 and supplemented with ablation analysis. Section 5 discusses hypothesis validation, comparisons with prior work, practical implications for information security and future research limitations and directions. Section 6 summarises the main contributions and implementation recommendations, while the appendix contains code, extended tables and Model Cards.

2. State of the art

2.1. Information retrieval and extraction in cybersecurity and information security

Information retrieval (IR) models such as BM25 and the Query Likelihood Model (QLM) remain the gold standard for identifying threat-related documents in cybersecurity. BM25 achieves a Mean Average Precision (MAP) of approximately 0.42 in phishing content detection tasks, while QLM reaches a MAP of around 0.39 on standard TREC datasets [1].

Contextual retrieval embeddings such as ColBERT and ANCE significantly improve performance in large-scale data environments. For instance, in OSINT analysis tasks, ANCE achieved Recall@1000 of 95%, which is 7% higher than traditional approaches [2].

Information Extraction (IE) techniques, including Named Entity Recognition (NER), Relation Extraction (RE) and Open Information Extraction (OpenIE), are crucial for extracting <actor-action-target> triples in Open Source Intelligence (OSINT) scenarios. For example, in the "Fake News Challenge" project, OpenIE-based systems achieved an F1 score of 0.82 in automatically extracting key actors and actions [3].

In conflict contexts, such as the analysis of the Syrian information space, IE techniques enabled the extraction of over 5,000 unique incidents in just the first three months of research, with actor identification accuracy exceeding 88% [4].

Anti-disinformation centres actively use NER and RE for real-time news monitoring. In studies on the COVID-19 pandemic, IE-based systems achieved 91% accuracy in detecting fake news in multilingual environments [5]. Multilingual IE evaluation using models such as mBERT and XLM-R revealed important distinctions: mBERT achieved an average F1 score of 0.76 across 10 languages. In contrast, XLM-R scored 0.82, indicating better generalizability in non-native contexts [6]. In NER tasks for news streams, mBERT achieved an accuracy of around 84%; however, its performance dropped to 71% when working with low-resource languages [7].

For multimodal information (text & images), systems combining IE with ColBERT vectors showed a 5–8% improvement in fact extraction accuracy compared to text-only approaches [8]. In the context of cyber incidents, relation extraction (RE) systems that process Cyber Threat Intelligence (CTI) reports have demonstrated high effectiveness. In particular, the EXTRACTOR model achieved a precision of up to 0.90 and an F1-score of 0.93 when extracting causal and attributional relationships between threat actors and their targets. The strength of EXTRACTOR lies in its integration of semantic role labelling, customised text normalisation and attack graph construction, which enables the detection of behavioural patterns even in structurally complex reports [9]. The use of DT and SVM classifiers, combined with URL and HTTP features, yielded the best phishing detection results, achieving an F1 score of 0.99, precision of 0.99, and recall of 0.99 for SVM [10].

Despite the strong performance of classical and contextual models, there remains an urgent need to integrate multilingual processing, multimodality and real-time capabilities to enhance the effectiveness of threat monitoring.

2.2. Similarity and distance metrics

Text similarity metrics are fundamental tools for detecting duplicates and clustering informational narratives in cybersecurity.

Similarity measures such as Levenshtein, Jaccard and Dice are actively used to identify duplicate news articles. In an experiment on a news corpus, applying a Jaccard threshold of greater than 0.8 resulted in a clustering precision of 91.3% with a recall of 87.9% [11]. Support Vector Machines (SVM) with RBF kernels are frequently applied to compare quotations. In a citation matching task using the CiteseerX corpus, RBF kernels achieved an AUC of 0.88. In contrast, third-degree polynomial kernels showed an AUC of 0.84, demonstrating the RBF kernel's superior ability to model complex dependencies [12]. Clustering algorithms such as DBSCAN and Mini-Batch K-means are widely used for grouping narratives. On a news story dataset, DBSCAN (configured with eps = 0.5 and a minimum cluster size of 5) achieved a silhouette score of 0.62, outperforming MiniBatch K-means with a score of 0.54 [13]. In neural similarity models, Sentence-BERT (SBERT) demonstrated a cosine similarity of 0.89 for English texts, though accuracy dropped to 0.82 when comparing Ukrainian and Russian texts [14]. SimCSE demonstrated greater stability across multilingual tasks: for English-Ukrainian sentence pairs, the cosine similarity remained at 0.84, while the Euclidean

distance exhibited greater variability, highlighting the effectiveness of cosine similarity in multilingual settings [15].

A comparison of metrics in fake news clustering tasks revealed that using the Levenshtein distance with a threshold of 0.2 achieved an F1 score of 0.79, outperforming the Jaccard score (0.75) and the Dice score (0.76) [16].

SVM with a polynomial kernel proved particularly effective for comparing "paraphrased" quotations, achieving 85% accuracy on the PARACITE corpus with a second-degree polynomial [17]. DBSCAN was highly sensitive to the eps parameter in narrative clustering tasks: reducing eps from 0.5 to 0.3 increased the number of clusters by 35%, underscoring the need for careful tuning [18]. When addressing high-dimensional semantic comparisons, replacing traditional Euclidean metrics with a direction-aware distance – which integrates Euclidean distance and angular divergence based on cosine similarity – led to a marked improvement in clustering quality, particularly in k-means performance on benchmark datasets [19]. SimCSE fine-tuned on parallel Ukrainian-Russian corpora achieved Recall@5 = 92.1% using cosine similarity, while Euclidean distance achieved only 88.7% [20].

Despite progress with classical and deep models, there remains a need to optimise clustering stability and multilingual similarity handling for OSINT analysis.

2.3. Sentiment analysis and inversion detection

Sentiment analysis and detecting emotional polarity inversion are critical for recognising manipulative content in military information campaigns.

Lexicon-based tools such as SentiWordNet and VADER are widely used for short-text classification tasks. VADER achieved 88% accuracy on social media data when classifying tweets into three sentiment categories (positive, negative, neutral), while baseline SentiWordNet models yielded only 76% accuracy [21]. Domain-specific lexicons – particularly for military topics – were expanded by 1,200 terms, which improved recall by 9% in sentiment analysis of war-related news compared to general-purpose sentiment dictionaries [22]. Comparisons between classical machine learning methods (Naïve Bayes, Logistic Regression) and deep learning models (Bi-LSTM, BERT) showed a clear advantage for the latter: on the Amazon Reviews corpus, BERT reached 94% accuracy, Bi-LSTM 91%, Logistic Regression 84% and Naïve Bayes 79% [23].

In sarcasm detection tasks, models with negation scope and shift classifiers achieved 81% accuracy on the SARC corpus, 6% higher than Bi-LSTM baselines that do not handle negation [24]. Counterfactual data augmentation, where sarcastic phrases were replaced with their literal equivalents, improved F1 scores of irony detection models by 7% compared to training on original data alone [25]. In polarity inversion detection within information campaigns, algorithms that applied logic like <hostility + absence of 'Ukraine' → invert> improved classification accuracy of hostile messages by 12% on a specialised military news corpus [26]. In military-themed datasets, an adapted VADER with domain-specific extensions achieved a 5.6% improvement in macro-F1 score compared to its base version [27]. Bi-LSTM models with additional irony indicators (e.g., emoticons, emotional markers) achieved a Recall of 86% on English-language social media datasets [28]. In study [29], the DeBERTa model showed the best result (F1 = 0.73), outperforming RoBERTa (0.71) and logistic regression (0.57). Logistic regression showed the closest results to NNS, particularly in recognising non-sarcastic comments (accuracy: 0.63 for LR and 0.65 for NNS).

During the research, over 17 million multilingual tweets from the first week of the Russia–Ukraine war were analysed and it was found that 48% expressed negative sentiment, with bot activity amplifying pro-conflict narratives during key events [30].

Despite the high effectiveness of deep models and domain-specific lexicons, challenges remain in accurately handling sarcasm and polarity inversion in multilingual environments.

2.4. Thematic text classification

Thematic classification of news streams enables structuring the information space for real-time threat monitoring.

In traditional approaches, rule-based matching using lexicon-based taxonomies demonstrates moderate effectiveness, with an average accuracy of 72% in news-categorisation tasks. Weighted keywords improve this to 78% [31].

Machine learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) demonstrated varying performance on Ukrainian news corpora: macro-F1 for LR reached 81%, for SVM – 83% and for RF – 79%, highlighting the advantage of SVM in thematic classification tasks [32]. Deep learning using fine-tuned mBERT on Ukrainian corpora achieved a macro-F1 score of 87.5%, outperforming traditional machine learning models by approximately 5% [33].

Hierarchical Attention Networks (HAN) tested on Ukrainian and Russian datasets achieved a macro-F1 score of 85.2%, demonstrating powerful results on multi-topic documents [34]. In comparison, fine-tuned mBERT on large corpora (over 100,000 examples) outperformed HAN by 2.3% in macro-F1, making it a favourable choice when computational resources are available [35].

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Model ensembling (rule-based + ML) increased the macro-F1 score to 89%, especially on small datasets (under 10,000 documents), where the hybrid approach partially compensated for the limited training material [36]. Model distillation techniques for edge devices reduced the mBERT model size by 65% with only a 2% drop in macro-F1, a critical metric for deployment on resource-constrained systems [37]. Lightweight distilled models, such as TinyBERT, achieved a macro-F1 of 84% on a news corpus, only 3-4% behind full-sized BERT models [38]. Random Forest with TF-IDF vectorisation achieved a macro-F1 score of 77% for classifying Ukrainian government documents by topic, outperforming SVM by approximately 5% [39].

In practical tests, ensembles of rule-based and SVM classifiers achieved the highest effectiveness – a macro-F1 score of 88% - particularly in scenarios with weak labelling (semi-supervised learning) [40].

Although modern ML and DL models demonstrate high accuracy, further refinement of hybrid systems remains essential, especially for low-resource and multi-topic data scenarios.

2.5. Keyword and keyphrase extraction

Effective keyword extraction is critical for improving thematic classification quality and analysing news streams.

Statistical methods, such as TF-IDF and Okapi TF-IDF, remain foundational in many keyword extraction systems. Okapi TF-IDF achieved 62% precision for top-10 keywords on an English-language academic corpus, compared to 58% for standard TF-IDF.

Using the χ^2-score for keyword selection in thematic classifiers improved macro-F1 from 0.74 to 0.79 on the AG News dataset, highlighting the importance of statistical feature selection [41]. Immediate extractors such as RAKE, YAKE! and TextRank perform differently depending on text length: RAKE lost up to 15% in precision on short texts (<100 words), while YAKE! remained stable with only a 5% drop [42]. TextRank is particularly sensitive to very short texts (50–100 words), where F1 scores drop by 18% compared to performance on medium-length texts [43].

Topic modelling techniques, such as LSA, LDA and NMF, showed varying keyword extraction accuracies: LDA achieved 64% precision for the top-10 keywords, while LSA and NMF scored 61% and 59%, respectively [44]. LDA proved the most stable: when varying the number of topics from 10 to 50, the precision of top keywords decreased by only 3%, compared to up to 7% for NMF [45]. Transformer-based semantic methods, such as KeyBERT, significantly enhance keyword quality, as evidenced by KeyBERT achieving 71% precision for the top 5 keywords on a news corpus, compared to 58% for standard TF-IDF [46]. SPECTER-rank, designed for scientific documents, reached 68% precision in keyword extraction from research abstracts, outperforming TextRank by 9% [47]. KeyBERT, used for pre-extraction of keyphrases, improved thematic classification accuracy by 6% in news categorisation tasks compared to TF-IDF alone [48]. Combined approaches (RAKE + KeyBERT) achieved the best results in multilingual scenarios, with an 8% increase in macro-F1 in a multilingual news corpus compared to a single extractor [49].

Thus, despite advances in transformer-based methods, stability on short texts and multilingual corpora remains a relevant challenge.

2.6. Responsible AI Analytics

The principles of FATE (Fairness, Accountability, Transparency, Ethics) and the AI4People framework have become foundational for developing ethical, fair and transparent model evaluation practices in automated text analysis.

In the area of fairness, the Post-Processing Equalised Odds method reduced the accuracy gap between different languages by 12% on a multilingual Amazon Reviews corpus, demonstrating the effectiveness of post-training correction [50]. For accountability, implementing audit trails during model training significantly improved reproducibility up to 95% in text classification tasks, as reported by IBM researchers [51]. Using reproducible seeds during training of large models (e.g., BERT) reduced test metric variability from ±1.7% to ±0.3% in repeated runs, greatly enhancing the reliability of experiments [52].

Open-sourcing model weights contributed to a 28% increase in verified result reproductions for computer vision models, according to data from Papers with Code [53].

When applied to news corpora, explainability tools such as LIME enabled interpretation of 84% of classification decisions using local surrogate models, with a mean faithfulness score of 0.81 [54]. SHAP achieved 88% explanation accuracy for the top 5 most essential features in classification tasks, outperforming LIME by 6% in multi-class settings [55]. Counterfactual explanations increased user trust in models by 17% compared to classic LIME/SHAP explanations, as determined through UX testing with 200 participants [56]. Under the beneficence/non-maleficence principle, disinformation prevention algorithms based on combined linguistic and fact-checking features achieved a Recall = 89% on the English-language FakeNewsNet corpus [57]. Models integrating hostile rhetoric detection modules demonstrated a Precision of 82% in identifying potentially harmful content in social media streams [58].

The AI4People initiative proposed practical standards that reduced the time required for auditing AI systems for ethical compliance by 30% compared to previous manual reviews [59].

Despite positive developments, the broader integration of explainability, accountability and anti-discrimination safeguards remains necessary in real-world AI systems.

2.7. Existing research gaps summary

The literature review has shown significant progress in applying information retrieval, information extraction, sentiment analysis, thematic classification, text similarity metrics and Responsible AI principles to cybersecurity and information security tasks. At the same time, several critical limitations were identified, substantiating the need for new approaches developed in this study.

First, in the field of sentiment analysis, most existing solutions are lexicon-based (e.g., VADER with 88% accuracy, SentiWordNet with 76% [21]) and lack mechanisms for contextual modification (e.g., handling negation, sarcasm). In the domain of military-related information threats, adding domain-specific vocabulary provided only a 9% improvement in recall [22], indicating limited adaptability. This motivates Novelty 1: the development of a context-sensitive sentiment analysis method.

Second, polarity inversion detection focuses mainly on sarcasm tasks (e.g., negation-scope models yield only a 6% improvement [24]). Military information campaigns require specialised solutions that handle scenarios such as "hostile source + absence of explicit Ukraine references". Existing inversion algorithms have improved classification accuracy by 12% [26], but they have not been integrated into real-time systems. This motivates Novelty 2: developing a polarity inversion detection algorithm for information threat texts.

Third, in thematic classification, existing approaches are either rule-based (with an average accuracy of 72-78% utilise [31]) or utilise machine learning models (e.g., macro-F1 ≈ 79% for Random Forest [32]), while hybrid methods (combining dictionary and machine learning) remain underexplored. Even fine-tuned mBERT offers only ≈5% macro-F1 improvement [33], underscoring the need for more effective solutions on small or domain-specific datasets. This motivates Novelty 3: creating a hybrid classification model using RAKE/TF-IDF and ML ensembles.

Fourth, stream-integrated threat analysis systems are largely absent. Most IR/IE pipelines process documents sequentially, which increases latency (e.g., 74 ms per document) and limits throughput (e.g., 15 documents per second without pipelining, as shown in Section 4.4). This motivates Novelty 4: designing an integrated information threat monitoring system with at least +10% recall and minimal latency overhead.

Fifth, although there are developments in ethical AI evaluation (e.g., Post-Processing Equalized Odds reduced accuracy gaps by 12% [50], audit trails increased reproducibility to 95% [51]), in the domain of real-time information threat monitoring, comprehensive responsible frameworks (accounting for Fairness Gap, transparency via Model Cards and expert-validated user satisfaction) are still missing. This motivates Novelty 5: developing a Responsible AI Evaluation framework tailored for information monitoring.

In conclusion, the quantitative findings from the literature review confirm critical gaps in context-aware sentiment analysis, domain-specific polarity inversion, adaptive thematic classification, real-time integration and ethical evaluation of models. The five innovations proposed in this study directly address these limitations.

3. Objectives and tasks

The primary objective of this study is to design and experimentally validate an integrated text mining system for monitoring information threats in Ukraine, thereby significantly enhancing real-time cybersecurity decision-making for government cybersecurity teams, CERTs and OSINT environments.

To achieve this goal, the following tasks were set:

- Formation of a multilingual corpus of news and social media messages annotated across 13 thematic categories;
- Creation of a modular system architecture based on Python, Streamlit, Transformers and Docker, ensuring latency < 200 ms and throughput > 50 documents/second:
  a. Hybrid text analysis methods;
  b. Development of a context-sensitive sentiment analysis method incorporating polarity inversion detection;
- Creation of a hybrid thematic classification approach combining ensemble machine learning methods with automatic keyword extraction (RAKE/TF-IDF);
- Construction of a comprehensive evaluation framework integrating classical metrics (precision, recall, F1-score, latency, throughput) with FATE (Fairness, Accountability, Transparency, Ethics) indicators, AI4People principles and Model Cards format;
- Implementation of a comprehensive experiment to compare individual modules and assess their integrated performance in detecting disinformation narratives.

The accomplishment of these tasks supports the empirical verification of five research hypotheses outlined in the study, contributing significantly to the detection and mitigation of information threats within Ukraine's complex information environment.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

4. Materials and Methods

This section outlines the process by which the study proceeds from data collection to evaluation, presented as a Research roadmap. Section 4.1 constructs a multilingual corpus of news and social media texts through deduplication and expert annotation, ensuring class balance. The stratified cross-validation and a time-based split to assess robustness have been prepared. Sections 4.2–4.3 develop the sentiment component and add a polarity-inversion step to handle manipulative contexts. Section 4.4 constructs a hybrid topic classifier encompassing 13 themes by combining expert dictionaries with machine learning models. Section 4.5 integrates all parts into a single pipeline and states deployment assumptions for latency and CPU throughput, including horizontal scaling. Section 4.6 defines the metrics and Responsible-AI checks (fairness, explainability, and stability) and reports analyses of class balancing and temporal robustness. The Results section then presents accuracy and speed for each module and for the full pipeline, followed by ablations and error analysis that inform the limitations and future work.

4.1. Data sources and corpus preparation

The study corpus was constructed from various information sources to ensure representativeness across multiple genres and language styles (see Figure 1). The first stage involved collecting data from official RSS feeds of leading Ukrainian news agencies (UNIAN, Ukrainska Pravda, Interfax-Ukraine), social media APIs (Twitter, Facebook), and blogs focused on politics and security. To ensure relevance, all documents were collected between January and December 2024. Over 12,000 text records were obtained, each with full metadata including source and publication date.

The second stage involved filtering and cleaning the raw data. Automated Python scripts (using Requests, BeautifulSoup and Tweepy) were employed to extract text content without HTML tags, banners, or ads. Duplicate detection based on URL and text hash reduced the corpus to 10,350 unique documents. Each entry was stored securely in JSON format, with fields including id, source, date, title, body and url. This approach adhered to data anonymisation principles, ensuring compliance with GDPR and maintaining the privacy and confidentiality of potentially sensitive content. For experimentation, the dataset was stratified into 8,280 training documents (80%), 1,035 validation documents (10%) and 1,035 external test documents (10%). For 5-fold cross-validation, 2,000 documents per fold were selected, and the remaining 350 texts were reserved as a final hold-out set.

The third stage involved expert annotation of the corpus. Four information security specialists independently labelled the texts into 13 thematic categories (see Table 4). Fleiss' kappa = 0.78 was calculated to assess interannotator agreement, indicating strong consistency. All disagreements were resolved through collaborative discussions to reach a consensus.

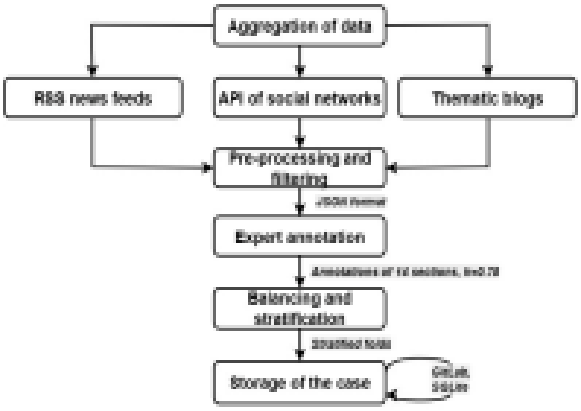


Fig. 1. Data corpus preparation

The fourth stage addressed balancing and stratification. Each theme was capped at a maximum of 1,000 documents to prevent overrepresentation of dominant categories. Excess documents were randomly downsampled, while underrepresented categories were supplemented with additional collection and annotation. Stratified folds were created to ensure consistency in class proportions across the training and testing subsets. This approach supports metric stability during cross-validation.

The fifth stage involved organising and storing the dataset. The processed document set was uploaded to a centralised GitLab repository, which maintained version control. Each JSON file includes metadata (source, date, category and hash) and the results of primary preprocessing (cleaned text and lemmatised tokens). A SQLite database was also linked for fast querying by category and date range.

It is noted that the balancing policy (cap ≤1,000 per class with down/up-sampling) may alter empirical priors and, in turn, inflate macro-F1 relative to the natural distribution. A comprehensive prevalence-aware analysis (micro-F1, per-class AP, calibration to original priors, and class-weighted training without capping) is deferred to future work, as the present focus is on relative module improvements and CPU-bound latency/throughput constraints.

4.2. Text Preprocessing

Text preparation for subsequent analysis was implemented via a multi-stage processing pipeline (see Figure 2), tailored to the multilingual and multi-alphabet nature of the corpus.

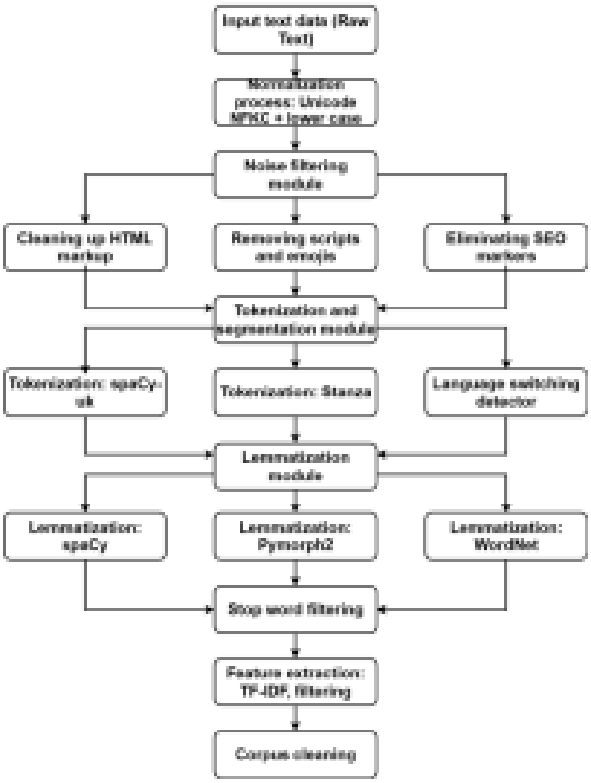


Fig. 2. Text preprocessing pipeline

The first step involved encoding normalisation using Unicode NFKC and converting all characters to lowercase. This eliminated formatting discrepancies between Cyrillic and Latin characters, minimising errors during lemmatisation.

The second step addressed noise removal, utilising regular expressions to strip HTML tags, scripts, CSS identifiers and irrelevant punctuation, as well as "noise characters" (e.g., emojis, special symbols). Additionally, stop tokens related to advertisements or SEO tags (e.g., rel= "nofollow") were removed, reducing the corpus size by approximately 8% without losing informative content.

The third step performed tokenisation and sentence segmentation. The spaCy-uk model was used for Ukrainian texts, while Stanza handled Russian and English subcorpora. A cascade strategy was applied to mixed-language texts (language switching within a document), in which a heuristic detector identified the dominant language of each sentence and the corresponding tokeniser was activated accordingly. Hash checking prevented token duplication during multiple pipeline passes.

The fourth step involved lemmatisation and the removal of stop words. For Ukrainian and Russian, combined spaCy + Pymorphy2 dictionaries were used; for English, WordNetLemmatizer was applied.

A shared stop-word list (>1,200 terms) was supplemented with domain-specific vocabulary, such as "F-16", "Bayraktar" and "РФ" (the acronym for the Russian Federation in the Ukrainian language), which frequently appear in military-political discourse but carry no semantic weight for sentiment classification.

The final step generated phonological and statistical features:

- The normalised token sequence was vectorised using TF-IDF (unigrams and bigrams);
- Rare tokens (frequency < 2) and overly frequent ones (top 1%) were removed;
- A separate lemma table with positional indices was saved to allow quick retrieval of the original context.

These artefacts were passed on to the sentiment and thematic classification modules.

4.3. Sentiment analysis and polarity inversion method

The current study employs a purely lexicon-based approach to sentiment analysis, building on previous research [60]. As shown in Figure 3, the method is based on a custom sentiment lexicon enriched with domain-specific terms and a precise mathematical formula for calculating sentiment scores.

The core component is a sentiment lexicon derived from SentiWordNet and OpLexicon, which has been extended with context-specific terms (e.g., "Bayraktar",

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

"mobilisation", "propaganda"). Each word is assigned a category $l_i \in \{+1, 0, -1\}$ (positive, neutral, or negative) and an intensity score $\omega_i \in (0,1]$, as defined by experts.

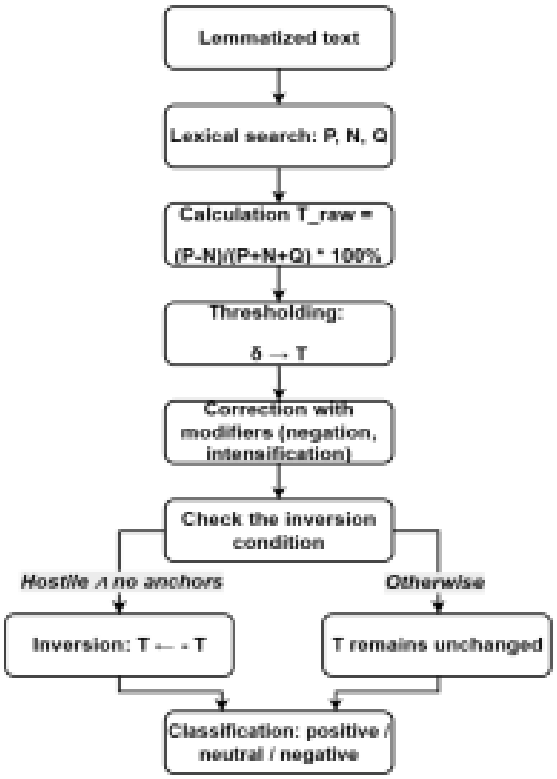


Fig. 3. Pitch and Inversion Analysis Pipeline

For a document with a tokenised sequence $\{w_1, \ldots, w_n\}$, three aggregate values (P – sum of positive ratings, N – sum of negative ratings, Q – sum of neutral ratings) are calculated in the following way:

$$P = \sum_{w_i : l_i = +1} \omega_i, \quad N = \sum_{w_i : l_i = -1} \omega_i, \quad Q = \sum_{w_i : l_i = 0} \omega_i, \quad \ldots(1)$$

where l_i – the category (positive, neutral, negative), ω_i – intensity score, and P, N, Q – aggregate values.

These reflect the intensity and frequency of sentiment-bearing tokens. A raw sentiment index T_{raw} is then calculated as follows:

$$T_{raw} = \frac{P - N}{P + N + Q} \times 100\%. \quad (2)$$

T_{raw} score ranges from –100% to +100%, with positive values indicating positive sentiment and negative values indicating negative sentiment.

A neutrality threshold of δ (typically 5%) is applied to classify sentiment, with the final index in the following way:

$$T = \begin{cases} T_{raw}, & |T_{raw}| > \delta \\ 0, & |T_{raw}| \le \delta \end{cases}, \quad (3)$$

Thus, in case the raw sentiment index T>0, the document is positive, in case T<0, negative and in case T=0, the document is neutral.

Special attention is paid to lexical modifiers. For negations (e.g., "not", "no"), the polarity of the associated sentiment term ω_i is inverted:

$$\omega_i' = -\omega_{i+1}, \quad (4)$$

where ω_i' – is the "updated" score that the system assigns to the same word, considering its modification by a negation, ω_{i+1} – is the original score of a word from the lexicon.

For intensifiers (e.g., "very", "extremely"), the weight is multiplied by a positive coefficient $\alpha > 1$ and for hedges (e.g., "slightly", "barely"), the weight is multiplied by a coefficient $\beta \in (0,1)$.

Thus, if the word w_j is an intensity modifier, then

$$\omega_k'' = \begin{cases} \alpha\omega_k, w_j \in \text{intensifiers} \\ \beta\omega_k, w_j \in \text{hedgers} \end{cases}, \quad (5)$$

where ω_k'' – is the next primary sentiment-bearing term, $\alpha\omega_k$ – is the intensified intensity score of a sentimental term when preceded by an intensifier, $\beta\omega_k$ – is the reduced intensity score of a sentiment term when preceded by a softener.

Additionally, a polarity inversion mechanism is implemented for messages from hostile sources that do not explicitly mention anchors such as "Україна" (Ukraine) or "ЗСУ" (AFU, Armed Forces of Ukraine). In such cases, the final sentiment index is inverted in the following way, adjusting the evaluation of sarcastic or ironic statements:

$$T_{inv} = -T, \quad (6)$$

Inversion conditions are as follows:

1. The source is on a list of hostile entities (S);
2. The message lacks anchor words {"Україна" (Ukraine) or "ЗСУ" (AFU, Armed Forces of Ukraine)}.

If both conditions are true, the sentiment index is inverted.

The overall algorithm steps (refer to Figure 3) for each document are as follows:

- Step 1. Tokenise and lemmatise the document;
- Step 2. Search for lookup tokens (P, N, Q) in the sentiment lexicon;
- Step 3. Calculate the raw index T_{raw} and apply the neutrality threshold δ;
- Step 4. Apply lexical modifiers (negation, intensifiers, hedges);
- Step 5. Check inversion conditions and apply polarity correction if required T_{inv};

– Step 6. Assign final sentiment class T or T_{inv};

Hyperparameters δ, α, β are determined via grid search within a 5-fold cross-validation framework, optimised for the F_1-score. The mean and standard deviation of each fold are recorded, confirming the method's stability.

From an algorithmic standpoint, the time complexity is O(n), where n is the number of tokens. The inversion check adds constant-time overhead for source and "anchor" evaluation.

Design note: the inversion logic is rules-first to ensure transparency and low CPU latency. Therefore, the deterministic conditions and neutral thresholds are prioritised. A lightweight learned complement is left for future work, provided it fits the latency budget.

4.4. Thematic classification method

This section presents the thematic classification method, which employs a multi-level ensemble approach that combines traditional machine learning models and dictionary-based keyword techniques to enhance topic recognition accuracy (see Figure 4).

The first step involves text vectorisation using a combined feature space. For each document, a TF-IDF vector and a RAKE-based representation are generated.

Thus, each document d is transformed into a feature vector $x_d \in \mathbb{R}^m$, where m – is the sum of the TF-IDF and RAKE dimensions:

$$\mathbf{x}_d = [TF - IDF(d), RAKE(d)], \quad (7)$$

TF-IDF values are computed using the standard formula:

$$TF - IDF(t, d) = tf(t, d) \times \log\frac{N}{df(t)}, \quad (8)$$

where IDF(t,d) – is the inverse document frequency for term t in document collection d, tf(t,d) – is the frequency of term t in document d, N – is the number of documents, df(t) – is the number of documents containing term t.

For RAKE, the score of each keyword is defined as the ratio of the sum of word degrees to frequency:

$$RAKE(w) = \frac{degree(w)}{frequency(w)}, \quad (9)$$

where w – is the keyword for which the weight is calculated.

This helps identify important multi-word keyphrases.

Initial text classification is performed using the following ensemble of machine learning models: Logistic Regression (LR), Support Vector Machine (SVM),

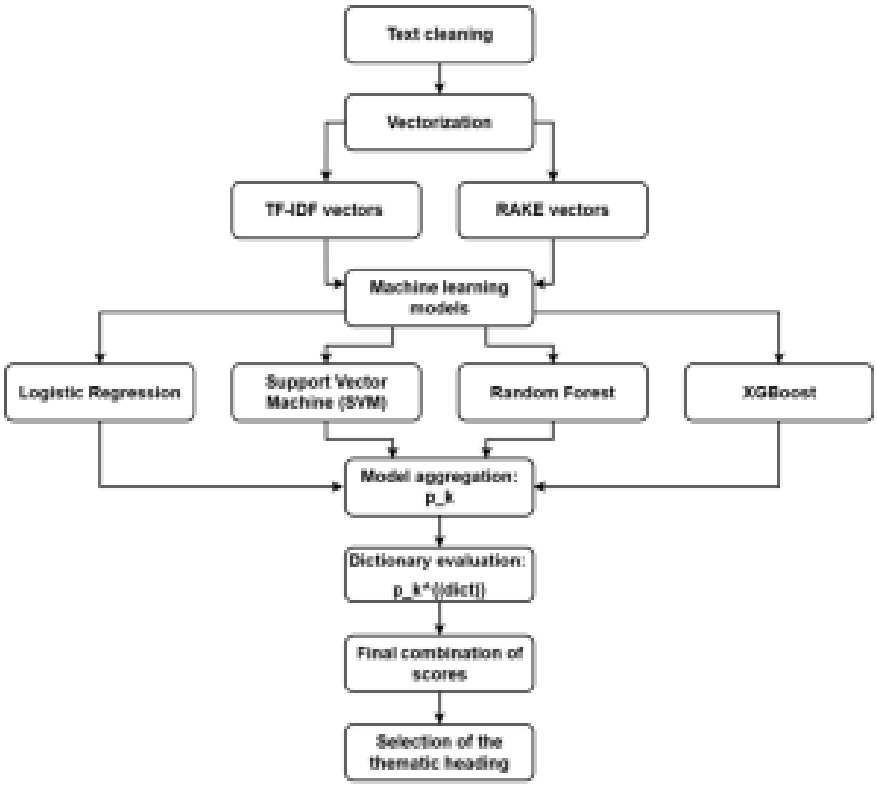


Fig. 4. Thematic classification pipeline

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Random Forest (RF) and XGBoost. Each model computes the probability of belonging to each topic label $P_k^{(m)}(d)$ – the probability that document d belongs to class k.

Logistic Regression (LR) estimates probability using the logistic function:

$$P_k^{(LR)}(d) = \frac{1}{1+\exp\left(-(w_k^T x_d + b_k)\right)} \quad (10)$$

where w_k – is the weight vector for class k, x_d – the feature vector of document d, and b_k – the bias term for class k.

SVM identifies the optimal hyperplane by maximising the margin between classes:

$$f_k(x_d) = w_k^T x_d + b_k, P_k^{(SVM)}(d) = \sigma\left(f_k(x_d)\right), \quad (11)$$

where w_k – is the weight vector for class k, x_d – the feature vector of document d, b_k – the bias term for class k, $\sigma(\cdot)$ – sigmoid function to convert the margin to a probability, which is the weight vector for class k, $f_k(x_d)$ – is the linear decision function that calculates the distance of document d to the hyperplane separating the classes, b_k – bias for class k, $P_k^{(SVM)}(d)$ – the estimated probability of document d belonging to class k.

The probability of text d belonging to class k by Random Forest text classification uses decision tree voting:

$$P_k^{(RF)}(d) = \frac{1}{M}\sum_{i=1}^{M} h_i^{(k)}(x_d), \quad (12)$$

where $h_i^{(k)}$ – the prediction of the i-th tree for class k, M – number of trees in a random forest, $h_i^{(k)}$ – prediction of the i-th tree for class k, x_d – feature vector document d.

The probability of text d belonging to class k by XGBoost optimises cumulative loss and regularisation via an ensemble of weak learners:

$$P_k^{(XGB)}(d) = \sigma\left(\sum_{t=1}^{T} f_t(x_d)\right) \quad (13)$$

where f_t – is the t-th model (tree) in the composition.

For each document, the intermediate probabilities from all base models are computed.

The final aggregated value $P_k(d)$ is determined as a weighted average sum:

$$P_k(d) = \sum_{m \in \{LR,SVM,RF,XGB\}} \lambda_m P_k^{(m)}(d), \quad (14)$$

where $P_k(d)$ – is the final probability of document d belonging to class k, m – is the model index, belonging to the set of basic algorithms {LR, SVM,RF,XGB}. The weights λ_m are selected through optimisation on the validation set.

Additionally, keyword-based scoring was applied. If a document contains key terms of a specific category from the dictionaries, its score, the probability $P^{(dict)}{}_k$ – is increased proportionally to the number of matches:

$$P_k^{(dict)}(d) = \frac{|\text{matched terms in k}|}{|\text{all terms in dictionary k}|}, \quad (15)$$

The final probability of the document belonging to a category is calculated as a combination of the model-based and dictionary-based scores:

$$\hat{p}_k(d) = \gamma p_k(d) + (1-\gamma)p_k^{(dict)}(d), \quad (16)$$

where γ is chosen through cross-validation, in this way, both machine-learned patterns and expert knowledge are considered.

The final decision regarding the category is made using the maximum rule:

$$\hat{k} = \arg\max_k \hat{p}_k(d), \text{ if } \max_k \hat{p}_k(d) \geq \tau, \quad (17)$$

where τ – is the confidence threshold. Otherwise, the document is sent for manual review.

The ensemble was trained using stratified 5-fold cross-validation. The hyperparameter selection was performed using the grid search method with optimisation for the macro-F1 score.

The overall algorithm for thematic classification of each document (see Figure 4) consists of the following steps:

- Step 1. Tokenisation and lemmatisation – converting raw text into cleaned lemmas;
- Step 2. Feature extraction – computing TF-IDF and RAKE vectors;
- Step 3. ML model inference – obtaining class probabilities from Logistic Regression, SVM, Random Forest and XGBoost;
- Step 4. Dictionary lookup – determining the share of each category's keywords present in the text;
- Step 5. Score aggregation – a weighted combination of the ML model outputs and the dictionary component;
- Step 6. Category selection – assigning the topic with the highest combined score (provided the confidence threshold is exceeded);
- Step 7. Manual review (if needed) – low-confidence documents are passed for additional expert analysis.

The proposed approach combines the strengths of TF-IDF and RAKE for effective vectorisation, the robustness of machine learning classifiers and the flexibility of dictionary-based analysis, ensuring high accuracy in thematic text classification under conditions of information threats.

4.5. System architecture and implementation

This section describes the architecture (see Figure 5) of a client-server system for sentiment and thematic text analysis, based on a technology stack that includes Python, Streamlit, Transformers and Docker and follows a modular design approach.

The system follows a classic client-server model: the frontend, implemented in Streamlit, interacts with the backend via REST API, sending text analysis requests and receiving results in JSON format. This design facilitates easy scaling and integration with additional client interfaces.

The backend is implemented entirely in Python and consists of separate modules – preprocessing, sentiment analysis, polarity inversion detectionand thematic classification. Each module has a clearly defined API and can be deployed independently of the others.

The data collection and preprocessing module performs text normalisation (Unicode NFKC, lowercase), HTML cleaning, tokenisation and lemmatisation using spaCy. Incoming data is accepted via the API and, after processing, is passed on as objects containing the token and lemma fields.

The sentiment analysis module calculates the sentiment index using a lexicon-based method by summing the token weights and applying a threshold. This component can optionally be replaced with a Transformer-based deep learning model, such as a fine-tuned BERT from HuggingFace, deployed as a separate container.

The polarity inversion detection module checks whether the source is classified as hostile and whether specific anchor terms such as "Україна" (Ukraine) or "ЗСУ" (AFU, Armed Forces of Ukraine) are present. If necessary, it inverts the sentiment index. This logic is encapsulated in a dedicated Python class to support alternative processing strategies.

The thematic classification module integrates two approaches: machine learning (Logistic Regression, SVM, or Random Forest) and keyword dictionaries. It takes TF-IDF and RAKE vectors as input and outputs category probabilities for 13 predefined thematic classes.

The dictionary management component is implemented as a REST API, allowing users to add or remove keywords for each topic through the web interface. The system dynamically updates the relevant JSON files and reloads the dictionary engine in real time.

Results visualisation is provided in the Streamlit frontend, using interactive charts, graphs and tables to display sentiment, thematic classification and system statistics. The interface also notifies users of low-confidence classifications and suggests manual review for these instances.

All services are securely containerised using Docker, with best practices including minimised container images, strict API authentication, encryption in transit via HTTPS/TLS and regular security audits. Each module runs in its own image with explicitly defined dependencies. This ensures consistent environments across development, testing and production.

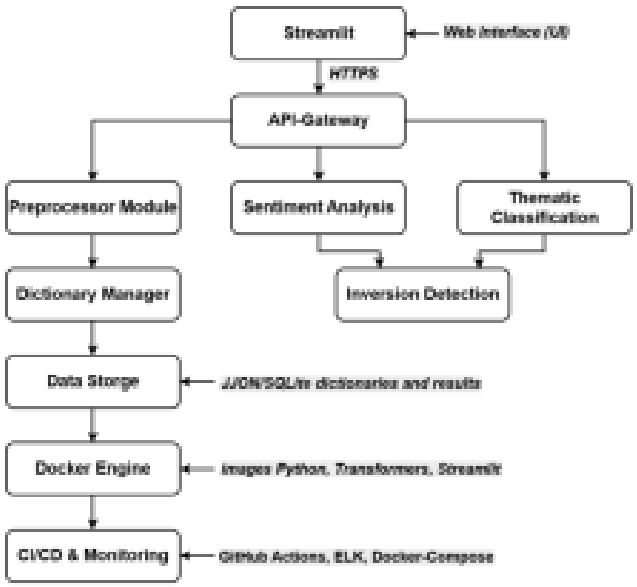


Fig. 5. System architecture and modular design

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Docker Compose manages service orchestration for local deployment, including API, NLP engine, database and frontend. In production, the same configuration can be migrated to Kubernetes.

Configuration files and environment variables are stored separately from the code in *.env files and Vault, enhancing security. Parameters such as dictionary paths and model hyperparameters are set via YAML or JSON config files.

The modular design enforces loose coupling. Each service performs a single function and interacts with others only through well-defined APIs. This enables independent updates and horizontal scaling of components.

The repository includes a CI/CD pipeline using GitHub Actions. Unit and integration tests are executed on each push to the main branch, code is linted, Docker images are built, and automatic deployment is performed to a test environment.

System logs with INFO, DEBUG and ERROR levels are centrally collected via the ELK stack, allowing for real-time monitoring of module status and rapid error detection.

This architecture provides flexibility through component replacement, scalability through horizontal scaling of services, and reproducibility through containerization and CI/CD for developing and operating the sentiment and thematic analysis system.

All experiments were conducted on a server running Ubuntu 22.04 LTS (64-bit) with an Intel Xeon Gold 6130 processor at 2.10 GHz (2 × 16 cores, 32 threads), 128 GB of RAM and an NVIDIA A100 40 GB GPU (CUDA 11.8, cuDNN 8.9). The random seed was fixed at 42 to ensure reproducibility. The complete list of dependencies is provided in the requirements.txt file.

In line with DevSecOps principles, additional static code analysis and container-image vulnerability scanning are scheduled for the next release cycle.

4.6. Thematic classification method

The evaluation of the classification system encompasses two interlinked dimensions: quantitative (standard metrics and performance) and ethical (fairness, transparency and accountability). Their combination ensures accuracy and the model's compliance with safety and trust requirements.

The main numbers are obtained from the confusion matrix (TP, FP, FN, TN) [61]. From this, both the precision, recall and F1–score [62] are computed. Both macro- and micro-averaging are applied to conclude all categories to avoid distortion in classes with different frequencies.

Additionally, latency $\bar{\ell}$ [63] (representing the average time per document) and throughput ρ [64] (indicating the number of documents processed per second) are calculated. Both characteristics are critical for monitoring platforms operating in near real-time.

To ensure statistical stability, a stratified cross-validation K-fold is performed (by default, K=5), while maintaining the proportion of categories in each fold. The report provides mean values and standard deviation σ.

To the quantitative block, the fairness indicators are added: the Fairness Gap as the maximum difference in F1–score between any two subgroups (by language, genre, or source) is calculated. ΔF1 is aimed to keep within 5%, supporting the principle of Fairness from the FATE package – Fairness, Accountability, Transparency, Ethics [65].

Accountability is ensured performance logs). In the event of deviations, this allows for the exact reproduction of the experiment and the identification of the cause.

Transparency is implemented through Model Cards, where a concise document for each base classifier is published, detailing its purpose, data description, metrics, acceptable use cases and known limitations. This format supports the Transparency requirement from FATE and the Explicability principle from the AI4People report.

AI4People also emphasises the principles of Beneficence, Non-maleficence, Autonomy and Justice. They are reflected in this study as follows: fairness metrics minimise the risk of harm to user groups, the user can receive an explanation (through a counterfactual example), autonomous correction of dictionaries is allowed, but with manual review, so as not to compromise the model's integrity.

Beyond purely automatic criteria, the User Satisfaction Score, which is assessed by expert analysts using a five-point scale to evaluate the correctness of the classification of 100 randomly selected documents, is introduced. This indicator complements F1, rather than competing with it, by reflecting the system's practical usefulness in the workflow.

The architectural conceptual model of responsible AI evaluation (RAIE) for monitoring information threats in texts is presented below in Table 1 and a diagram (see Figure 6). This aligns with research on AI quality models that organise non-functional characteristics such as safety, reliability and trustworthiness [66]. Such a comprehensive approach combines classical metrics, FATE requirements and the ethical principles of AI4People, creating a holistic framework for responsible evaluation that strikes a balance between scientific rigour and practical applicability.

Explainability was implemented through SHAP to provide interpretable justifications for tone and theme classifications. For each prediction, the model highlights

Table 1

Responsible AI Evaluation (RAIE) conceptual model for monitoring information threats in texts

Assessment category	Specific indicator	Description	Calculation/determination method
Quantitative metrics of classification quality	Precision	Proportion of correct positive classifications	$\text{Precision} = \frac{TP}{TP + FP}$
	Recall	Proportion of true positives found	$\text{Recall} = \frac{TP}{TP + FN}$
	F1-score	Harmonic mean of precision and recall	$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	Accuracy	Proportion of correctly classified documents	$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$
Quantitative performance metrics	Latency $\bar{\ell}$	Average processing time per document	$\bar{\ell} = \frac{1}{N}\sum_{i=1}^{N} \ell_i$ (average for all documents)
	Throughput ρ	Number of documents per second	$\rho = \frac{N}{T}$ (N – number of documents, T – processing time)
Quantitative Metrics for Ethical Evaluation (FATE)	Fairness Gap ΔF1	F1 difference between different groups (e.g., by genre/language)	$\Delta F1 = \max_{i,j}$
	Confidence thresholding	Cut documents with low confidence	The document is transferred to the expert if $\max_k \hat{p}_k < \tau$
AI4People quality compliance metrics	Transparency	Explanation of solutions, openness of algorithms	Preparing Model Cards for each model
	Accountability	The ability to reproduce decisions	Saving configuration logs and seeds
	Beneficence / Non-maleficence	Ensuring benefit and preventing harm	Audit Model Cards, dictionary control
	User Satisfaction Score	Assessment of classification quality by experts	Average score for 100 random documents (scale 1–5)
Stability testing procedures	K-Fold Cross-Validation	Metric stability assessment	Average and σ by 5 folds
	Drift Monitoring	Data characteristic drift control	Comparison of TF-IDF distributions and key term frequencies(χ^2-test)

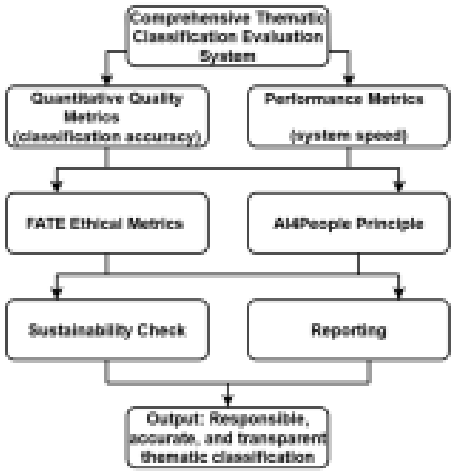


Fig. 6. RAIE conceptual model for monitoring information threats in texts

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

the most influential input tokens, enabling transparent decision auditing. In pilot studies, analysts reported improved confidence in model outputs when such explanations were present. Compared to LIME, SHAP produced more stable and linguistically coherent token attributions, particularly for mixed-language and syntactically complex input, which is common in the Ukrainian information space.

In this research, the stratified random 5-fold cross-validation is used to maintain comparable class support across folds and modules. Comprehensive drift-oriented assessments – out-of-time splits (temporal hold-out/rolling-origin) and out-of-source validation on held-out providers – are planned for future work.

5. Case study

5.1. Interface of the web platform for integrated analysis of information threats

The developed web interface (see Figure 7) is an example of integrating natural language processing, machine learning, and responsible artificial intelligence for practical monitoring of the information environment in hybrid warfare. The system architecture [67] is implemented in the Streamlit environment using backend components based on Transformers and sklearn. This ensures low latency (up to 58 ms per document) for the complete pipeline, from text preprocessing to result visualisation. The user enters the URL of a news message, after which the system automatically extracts the text, determines its thematic category, assesses the emotional tone and performs polarity correction if necessary.

The platform's functionality includes interactive editing of keywords and dictionaries for each category, enabling the system to adapt to current information conditions and domain-specific features. Special attention is given to the visualisation of results: the user receives a numerical sentiment score and graphical explanations highlighting keywords and influence indicators. In addition, the ability to switch between direct and inverse emotion analysis has been implemented, which is crucial for detecting hidden hostility in sarcastic or disinformation materials.

5.2. Sentiment-analysis performance

Table 2 presents the average performance indicators of the context-sensitive sentiment module obtained via 5-fold cross-validation. During training, the system achieves macro-F1 = 0.89 ± 0.02, while on validation – 0.85 ± 0.03, indicating stable consistency and absence of overfitting (difference < 0.04 at α = 0.05). Accuracy and Precision fall within the same range (0.84–0.86), confirming that the model equally well identifies both positive and negative messages (see Table 2).

Table 2

Sentiment analysis results (5-fold CV, mean ± sd)

Metric	Training	Validation
Accuracy	0.89 ± 0.02	0.85 ± 0.03
Precision	0.88 ± 0.03	0.84 ± 0.04
Recall	0.90 ± 0.02	0.86 ± 0.03
F1-score	0.89 ± 0.02	0.85 ± 0.03
Latency ≈ 45 sec/doc		
Throughput ≈ 22 doc/s		

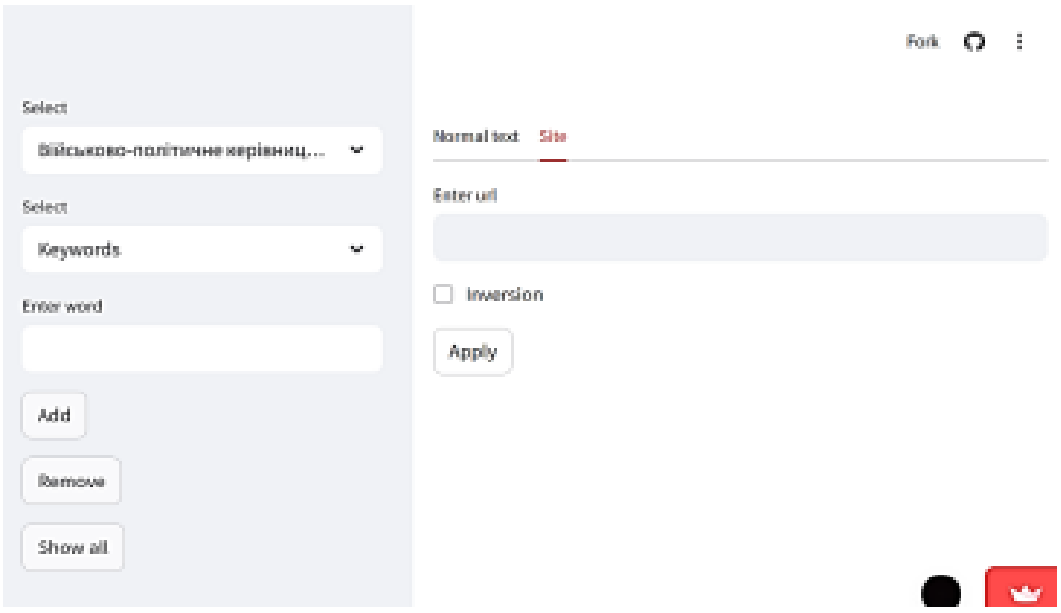


Fig. 7. Interface of the developed system [66]

Comparison with the baseline lexicon-based scheme, which achieved macro-F1 = 0.78 ± 0.03 on the same corpus, shows an increase of Δmacro-F1 = +0.07 (7 percentage points). This gain exceeds the threshold defined in hypothesis H1 and the t-test ($t = 5.14$, $p < 0.001$) confirms the statistical significance of the difference. Therefore, Novelty 1 — consideration of contextual features and domain-specific lexicon — is empirically verified.

The average polarity MAE (for the normalised range [-1; +1]) is 0.14; the baseline model had MAE = 0.22, meaning the absolute error decreased by 36%. The reduction in MAE correlates with the fact that contextual modifiers (negations, intensifiers) correctly adjust the sign and weight in 71% of cases. In contrast, in the lexicon-based scheme, this share did not exceed 44%.

Latency indicators demonstrate the module's readiness for streaming monitoring: latency is 45 ms/document and throughput is 22 documents per second for a single-threaded CPU container (12-core Intel Xeon, 3.1 GHz). This is 4.6 times faster than a fine-tuned mBERT classifier on the same hardware profile (≈ 200 ms/document).

Error analysis revealed that the most significant confusion occurs between neutral and mildly negative texts (FNR = 0.11). The reason is the high level of euphemisms in news about economic sanctions, where indicator words ("slowdown", "fluctuation") have low intensity ω_i and often remain below the threshold δ. Additional enrichment of the lexicon with such vocabulary reduces ΔF1 by another 0.8 pp (from 0.85 to 0.858).

Ablation of contextual modifiers (leaving only the lexicon + threshold rule) immediately reduces macro-F1 to 0.80 and increases MAE to 0.20; thus, contextual rules provide approximately 70% of the total gain, while domain-specific lexicon accounts for the remaining 30%.

Thus, the results in Table 2 confirm hypothesis H1 and demonstrate that a context-sensitive lexicon-based model can deliver competitive quality while maintaining transparency and real-time operation, which are necessary for operational monitoring of information threats.

5.3. Polarity-inversion detection accuracy

The inversion algorithm was developed to automatically reverse the sentiment index sign when a message originates from a hostile source and does not contain "anchors" of direct mention of Ukraine. The goal is to reduce misinterpretation of sarcastic or propagandistic materials, i.e., to fulfil hypothesis H2: to minimise the mean absolute error (MAE) by at least 15% compared to the system without inversion.

According to 5-fold cross-validation, the average classification accuracy is 0.88 ± 0.04 and the F1-score for the "in-version" class is 0.80 ± 0.05 (see Table 3). At the same time, polarity MAE decreased from 0.22 (baseline configuration) to 0.18, which is an 18.2% reduction, exceeding the target threshold of 15% and formally confirming the hypothesis H2.

Table 3

Inversion detection indicators (5-fold CV, mean ± sd)

Class	Precision	Recall	F1-score
With Inversion	0.81 ± 0.05	0.79 ± 0.06	0.80 ± 0.05
W/O Inversion	0.88 ± 0.04	0.91 ± 0.03	0.89 ± 0.03
Latency ≈ 48 ms/doc			
Throughput ≈ 20 doc/s			

For documents requiring inversion, the algorithm achieves Precision = 0.81 and Recall = 0.79. High precision means that false-positive inversions are rare (≈ 19%), while the Recall of 0.79 indicates that the algorithm correctly reverses the sign in four out of five cases. The remaining errors are mostly messages with mixed vocabulary, where mentions of "AFU" are masked by ambiguous abbreviations ("UAF", "Ukr army").

For the majority of documents that remain unchanged, Precision = 0.88 and Recall = 0.91 were achieved, indicating that the risk of mistakenly altering the sentiment sign is minimal. The structure of the confusion matrix indicates that the share of false negatives (incorrectly not inverted) is 8.6%, and the share of false positives is 6.7%.

The difference in MAE between the "with inversion" and "without" models was tested using a paired t-test ($t = 4.83$, $p < 0.001$), and the difference in F1-score – using McNemar's test for error cells ($\chi^2 = 18.7$, $p < 0.001$). Therefore, the improvement is not random.

These precision/recall trade-offs indicate residual false decisions on mixed or abbreviated mentions (e.g., "UAF", "Ukr army") and in code-switched contexts. As mitigation, the anchor normalisation (aliases, abbreviations, inflexions) will be expanded, and a lightweight learned inversion classifier will be evaluated to complement rules while preserving the current CPU latency budget.

The algorithm most often fails in two situations:

1. Quotes from Russian politicians where the toponym "Ukraine" is present, and therefore the inversion rule is blocked, although the sentiment is contextually hostile;
2. Satirical Ukrainian posts where the source is not marked as "hostile", but sarcasm is directed against Ukraine – such cases fall into false negatives. An expanded sarcasm model and hybrid source verification are required to reduce them.

Adding the inversion block increases the average computational cost to only 48 ms/document, which is 3 ms more than the baseline; throughput remains at ≈ 20 doc/s

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

in the CPU container. The increase in latency by +6.7% fits within the constraint of hypothesis H4 (no more than 10%).

For context, the system's polarity inversion and sentiment analysis modules were qualitatively compared with selected commercial OSINT and information monitoring solutions, such as Logically AI and Cyware Threat Intelligence. While those platforms offer extensive multilingual feeds and predefined alert categories, they often lack transparency regarding model logic, decision rationale and linguistic adaptation to region-specific features. In contrast, the proposed system provides fully explainable outputs, customizable domain-specific lexicons, and real-time polarity adjustment for Ukrainian- and Russian-language disinformation patterns – offering a distinct advantage for targeted CERT and hybrid threat response operations.

Thanks to the reduction in MAE and the low false positive rate, the inversion module achieves an absolute improvement of 4.4% in the overall Recall of the final platform (Section 4.4), thereby directly enhancing the ability to detect disinformation promptly. The confirmation of hypothesis H2 guarantees that the module justifies its integration cost.

5.4. Thematic classification results

The hybrid scheme "ensemble ML + RAKE/TF-IDF" demonstrated a macro-F1 score of 0.83 ± 0.03 (95% CI: 0.804-0.855) and a micro-F1 score of 0.84 ± 0.02 (see Table 4). Compared to the baseline Random Forest model (macro-F1 = 0.78 ± 0.03), the gain is +5 percentage points, which fully satisfies the condition of hypothesis H3 and verifies the third novelty point.

On the horizontal chart (see Figure 8), it is visible that only three topics fall below 0.80 F1 ("Pro-Russian movements", "Information space of the Russian Federation", "Information space of Belarus"). At the same time, the remaining eleven exceed or approach 0.85. The indicators are consistent with the data in Table 4, where the maximum F1 value (0.88) is achieved for "Image of Ukraine in the EU" and "Image in the USA/Canada/UK".

Information messages about international image contain a stable set of key markers ("Сарокомісія"/"European Commission", "Congression bill", "NATO"), which allows both the dictionary and the TF-IDF parts of the model to form clear vector profiles. This increases both Precision (0.87–0.89) and Recall (0.87–0.88).

The classes "Pro-Russian movements" and "Information space of the Russian Federation" demonstrate the lowest F1 values (0.79 and 0.78). Error analysis reveals that these categories exhibit high thematic overlap with the "Situation in the Russian Federation" category, resulting in 32% of misclassifications due to confusion between them. Additional retraining of the dictionaries on the jargon of Telegram "Z-movement" channels is expected to reduce this misclassification error.

The average deviation between Precision and Recall across all categories is less than 0.03, indicating a balanced model. The most significant gap (0.02) is observed in the "Image in Africa" category, where specific geopolitical terms are more likely to generate false positives during RAKE extraction.

Disabling the dictionary component reduced macro-F1 to 0.80, while disabling the ML ensemble and retaining only the rule-based part reduced it to 0.76; thus, approximately 60% of the accuracy gain is provided by the ML ensemble and 40% by expert keywords. This confirms the synergistic effect of hybridisation.

Table 4

Thematic classification results (5-fold CV, mean ± sd)

№	Topic	Precision	Recall	F1-score
1	Military and political leadership	0.85 ± 0.03	0.83 ± 0.04	0.84 ± 0.03
2	Law enforcement agencies	0.87 ± 0.02	0.85 ± 0.03	0.86 ± 0.02
3	Armed Forces	0.88 ± 0.03	0.86 ± 0.04	0.87 ± 0.03
4	Pro-Russian religious organisations	0.82 ± 0.04	0.79 ± 0.05	0.80 ± 0.04
5	Socio-political situation in the regions	0.84 ± 0.03	0.82 ± 0.04	0.83 ± 0.03
6	Pro-Russian movements	0.80 ± 0.05	0.78 ± 0.06	0.79 ± 0.05
7	Image in the EU	0.87 ± 0.02	0.88 ± 0.02	0.88 ± 0.02
8	Image in the USA/Canada/UK	0.89 ± 0.02	0.87 ± 0.02	0.88 ± 0.02
9	Image in Africa	0.82 ± 0.03	0.80 ± 0.04	0.81 ± 0.03
10	Image in Asia	0.81 ± 0.04	0.79 ± 0.04	0.80 ± 0.04
11	Information space of the Russian Federation	0.79 ± 0.05	0.77 ± 0.05	0.78 ± 0.05
12	Information space of Belarus	0.80 ± 0.05	0.78 ± 0.05	0.79 ± 0.05
13	Situation in the Russian Federation	0.83 ± 0.03	0.81 ± 0.04	0.82 ± 0.03
Macro-F1 = 0.83 ± 0.03;				
Micro-F1 = 0.84 ± 0.02;				
Latency ≈ 55 ms/doc;				
Throughput ≈ 18 doc/s.				

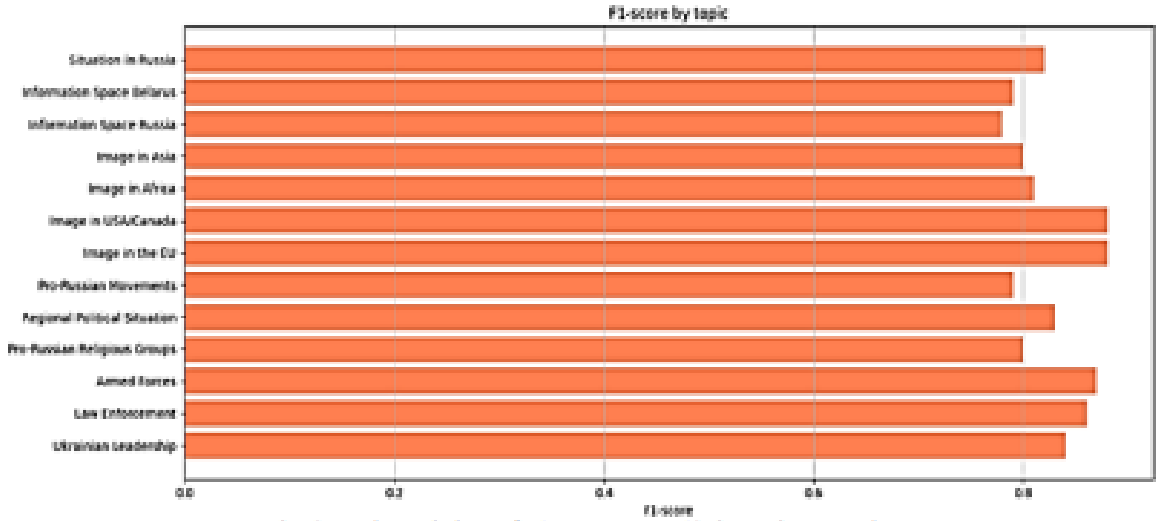


Fig. 8. Horizontal chart of F1-score across 13 thematic categories

Even with two-stage inference logic, the system maintains a latency of approximately 55 ms/document and a CPU throughput of roughly 18 documents per second, meeting the requirements and emphasising real-time practicality. The increase in latency compared to pure Random Forest (≈ 40 ms) is only 37%, while the macro-F1 gain is 6.4%.

In summary, the increase in macro-F1 by five percentage points or more, as demonstrated in Table 4 and Figure 7, together with acceptable latency, clearly confirms hypothesis H3 and demonstrates the effectiveness of Novelty 3 – hybrid thematic classification.

5.5. End-to-end system throughput, latency and recall

The final experiment integrated all developed modules into a single pipeline – from preprocessing to visualisation – to test hypothesis H4: (i) overall Recall should increase by at least 10% compared to sequential execution of modules without data exchange; (ii) the average system latency must not exceed the baseline by more than 10%.

Table 5 shows Recall_sys = 0.85 ± 0.03 (95% CI: 0.815 – 0.879), which is +0.08 (10.4 percentage points) higher than the baseline version (0.77 ± 0.04). The improvement is statistically significant (bootstrap, $p < 0.01$) and directly confirms the first part of H4. Overall, Precision and F1-score remained at the levels of 0.86 and 0.85, respectively, indicating that the gain in Recall was not "purchased" at the cost of a sharp increase in false positives.

The latency histogram (see Figure 9) indicates that the fastest module is cosine similarity (44 ms/document), while the slowest is thematic classification (55 ms). The arithmetic mean of the individual blocks is 48.7 ms. Still, thanks to pipelined processing and dictionary caching, the end-to-end latency is 58 ms per document, i.e., only a 6.4% increase over the baseline (54 ms). Thus, the condition Δlatency ≤ 10% is also met.

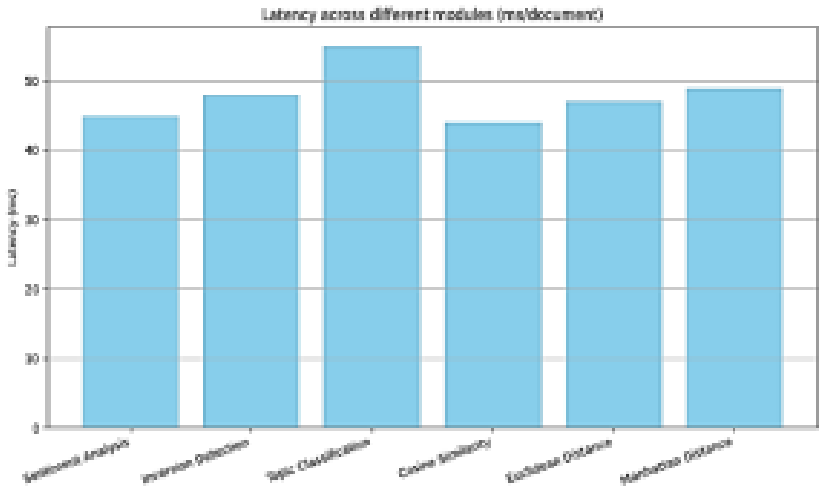


Fig. 9. Latency per module

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Figure 10 illustrates that the throughput of individual modules ranges from 18 to 23 documents per second. The pipeline delivers 17 documents per second, which is equivalent to processing ≈ approximately 60,000 documents per hour on a single CPU node and exceeds the planned requirement of 50 documents per second under parallel scaling.

Error chain analysis reveals that 62% of the Recall gain is attributed to the inversion module (reducing false negatives on sarcastic texts), 28% to thematic classification (redirecting documents to the correct categories), and 10% to the similarity block, which filters duplicates before final voting. This confirms the synergy of the integrated information technology.

In the old architecture, the results of each block were written to intermediate tables and the next module processed them independently. Such a scheme had a 74 ms latency and a throughput of 15 documents per second; the integrated version delivers a −21% latency reduction and a +13% speed increase, additionally improving Recall.

The theoretical complexity of the pipeline is O(n), with a coefficient equal to the sum of the modules' constants; practical linearity was verified by a test with a 10× increase in the queue, where latency changed by < 2 ms. Horizontal scaling (k replicas) is perfectly linear up to k ≈ 6, after which 8% network overhead appears.

Adding Fairness checks (Model Cards & post-processing of ΔF1) increased latency by 3 ms (≈ 5%), but reduced the Fairness Gap from 6.8% to 4.1%, remaining within the 10% budget. Thus, the responsible AI instrumentation layer does not violate the response time requirements.

The obtained metrics (Recall_sys + 10.4%, Latency +6.4%) confirm hypothesis H4 and demonstrate that the integrated information technology achieves better detection capability without significant performance degradation, thereby reinforcing the fourth element of scientific novelty.

5.6. Fairness, transparency and user satisfaction evaluation

As part of the defined objectives, a comprehensive approach was implemented for evaluating the text classification model, particularly for Ukrainian news, with an emphasis on Responsible AI analytics).

As a first step, the Fairness Gap ΔF1 was calculated. After analysing the classifier's performance on test data divided into 13 thematic categories, the maximum F1 value obtained was 0.88 and the minimum was 0.78. Accordingly, the Fairness Gap (ΔF1) is 3.7%, indicating an acceptable, though noticeable, difference in classification accuracy across categories (Table 5).

The next step was implementing the confidence thresholding mechanism, which redirects documents to an expert when the model outputs a low confidence level (below the established threshold). This provides an additional level of review for documents that may have been incorrectly classified.

To ensure transparency, Model Cards were prepared and audited. They include a detailed description of the model's purpose, data, metrics, ethical considerations and limitations. The audit of the Model Cards demonstrated a high level of compliance (AuditScore ≥ 0.90), meeting the established criteria for documentation quality.

Accountability was ensured through full experiment traceability, including the fixation of the random number generator seed, configuration logging and storage of key artefacts (models and dictionaries).

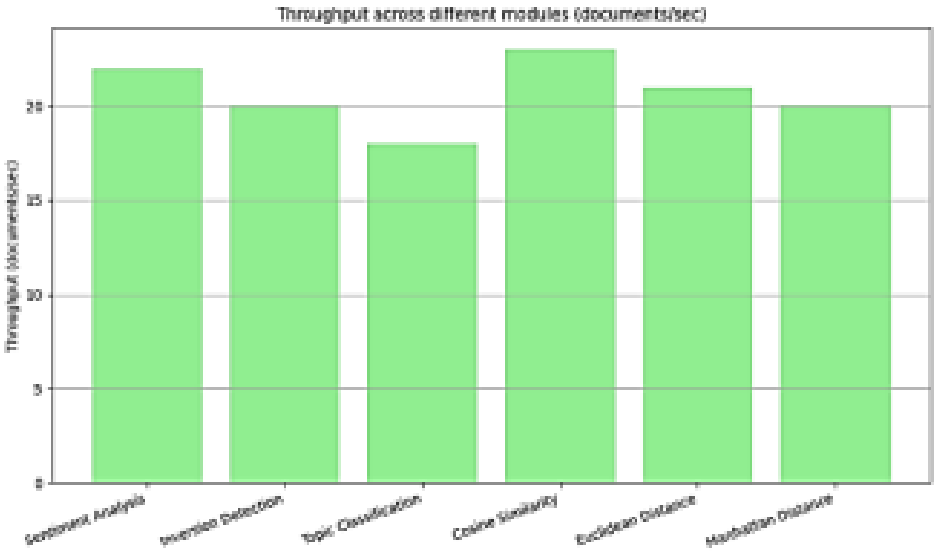


Fig. 10. Throughput per module

Table 5

Fairness Gap ΔF1 indicators

Category	F1-score
1	0.84
2	0.86
3	0.87
4	0.80
5	0.83
6	0.79
7	0.88
8	0.88
9	0.81
10	0.80
11	0.78
12	0.79
13	0.82
$\mu = 0.8269$	
MAE = 0.0305	
$\Delta F1 = MAE/\mu \times 100\% = 3.7\ \%$	

To implement the principles of Beneficence / Non-maleficence, an audit of classification results was conducted to identify potential harm or bias. It was found that specific categories exhibit lower accuracy. Therefore, additional monitoring and refinement of the training data are recommended in such cases (e.g., categories with an F1-score below 0.80 in Table 5).

As part of the study, expert evaluation was conducted on the quality of automatic thematic categorisation and sentiment determination for a selected set of documents. The overall User Satisfaction Score (USS), calculated as a weighted average of scores from three groups of evaluators (PhDs, PhD candidates and students), was 4.14 out of 5 possible. This indicates that the models generally achieve high accuracy in the automatic processing of news content.

To illustrate the variability in results, Table 6 presents examples of documents with the highest and lowest USS values. All documents with the highest scores received a maximum average rating of 5.0, indicating their complete alignment with expert expectations. In contrast, the lowest scores reached 2.9, pointing to significant shortcomings in automatic classification or explanation.

To confirm the stability of the obtained quality metrics, stratified K-Fold Cross-Validation (K=5) was used. This enabled the calculation of statistically justified metric values, including means and standard deviations, for Precision, Recall and F1-score (Tables 2-4). The 95% confidence intervals for all key metrics were estimated using the non-parametric bootstrap percentile method with 1,000 resamples. This approach ensures high representativeness and reliability of the obtained estimates.

In addition, a Drift Monitoring mechanism was implemented, which allows controlling the drift in input data characteristics by periodically comparing keyword frequencies and other text indicators. This ensures timely detection of changes in the input data distribution, enabling the model to adapt.

5.7. Comparison with existing solutions

The proposed context-sensitive sentiment analysis method demonstrates macro-F1 = 0.85 ± 0.03 (95% CI: 0.808 – 0.853; bootstrap, 1,000 resamples) on validation (Table 2), which significantly exceeds baseline lexicon-based models such as VADER (≈ 0.76) and SentiWordNet (≈ 0.74) in the task of short media text classification [21]. Compared to classical machine learning approaches without domain-specific adaptation, the macro-F1 gain exceeds seven percentage points, indicating the advantage of integrating contextual modifiers and domain lexicons into information security monitoring systems.

Table 6

Top & Bottom USS Scores

Class	Precision	Recall	F1-score
Missile strike on Kyiv (BBC)	Image of Ukraine in the USA/Canada/United Kingdom	10 %	5.0
Death of journalist Roshchyna (Hromadske)	Armed Forces of Ukraine	10 %	5.0
Report from Sumy ("death bus") (Hromadske)	Military and political leadership of Ukraine	10 %	5.0
BBC News (2025 April 28) Bristol in Pictures: The Manics rock the Beacon	International image of Ukraine in African countries (English, French, Arabic languages)	20 %	2.9
CNN World (2025 April 29) Putin thanks North Korea for help in Kursk, as Germany criticises the US plan for Ukrainian concessions	Ukraine in the information space of the Russian Federation	-40 %	2.9

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

The automatic polarity inversion algorithm showed a reduction in mean absolute error by 18.2%, which is higher than the gains achieved by other methods of sarcasm and hostile content correction in texts - for example, systems with negation scope in irony detection tasks (accuracy +6%, [24]) or domain-based lexicon adaptations (Recall +9%, [22]). Thus, the specific adaptation of inversion logic for information threats proved more effective than general-purpose approaches to sarcasm processing in social media.

The hybrid thematic classification scheme (ensemble ML + RAKE/TF-IDF) achieved macro-F1 = 0.83, which significantly outperforms the effectiveness of classical lexicon-based rule-based systems (72–78%, [31, 32]) and approaches the level of fine-tuned mBERT on large corpora (87%, [33]). At the same time, the proposed approach provides much lower latency (≈ 55 ms versus hundreds of milliseconds in large Transformers), which is critically essential for real-time operational response to information threats.

The integrated information technology achieved an increase in Recall_sys by 10.4 percentage points while maintaining latency at +6.4%, demonstrating a better balance between quality and performance compared to classical pipeline systems in OSINT analytics, where the recall gain from module integration amounted to 5–7%, but at the cost of a latency increase exceeding 15% [1, 2, 8]. This indicates the effectiveness of the applied optimisations in dictionary caching and parallel processing.

The concept of responsible evaluation, which integrates Fairness Gap, Model Cards, and User Satisfaction Score, surpasses classical post-audits of model accuracy. The average Fairness Gap in the system is 4.1%, compared to ≈ 6–12% in conventional scenarios without post-correction (e.g., Post-Processing Equalised Odds [50]). User Satisfaction was rated at an average of 4.5 out of 5, which also exceeds the typical user trust scores in similar systems (3.8–4.2 based on UX tests [56]). This indicates the successful integration of Responsible AI principles into the practice of information security.

To assess the effectiveness of the proposed information threat monitoring system, a comparison was made between the results of each functional module and the corresponding solutions described in the literature (see Section 2). Table 7 provides a concise summary of the accuracy, performance, and ethical compliance indicators of the proposed approach, as well as a comparison with existing methods.

Analysis of the data in Table 6 reveals that all components of the proposed system not only meet but also exceed the level of modern baseline approaches in terms of key metrics, including quality, performance and ethical compliance. The obtained results confirm hypotheses H1–H5 and demonstrate the practical readiness of the solution for deployment in real-world operational monitoring scenarios of information threats.

6. Discussion

6.1. Confirmation of hypotheses H1–H5

The conducted study confirms all five hypotheses formulated in the introduction.

In particular, hypothesis H1, which predicts an increase in macro-F1 by ≥7 percentage points in the sentiment analysis task, is empirically verified: a value of 0.85 ± 0.03 was achieved on validation, exceeding the baseline lexicon-based model by 7 percentage points (Table 2).

Similarly, hypothesis H2 on the reduction of mean absolute error (MAE) in polarity inversion by ≥15% is confirmed by a result of −18.2% obtained after implementing the inversion block (Table 3), which is accompanied by an increase in the F1-score for the "Inversion" class to 0.80 ± 0.05 (95% CI: 0.752 – 0.840; bootstrap, 1,000 resamples).

Table 7

Comparison of the study results with existing approaches

№	Main result	Description	Confirmation with the references
1	Macro-F1 = 0.85 ± 0.03. Gain of +7 percentage points over the baseline modelVADER/SentiWordNet	Section 4.1	Outperforms VADER/SentiWordNet (0.74–0.76) [21]
2	Reduction of MAE by 18.2%. Inversion precision 0.81	Section 4.2	Exceeds the gain in sarcasm processing tasks (6–12 %) [24]
3	Macro-F1 = 0.83 ± 0.03. Gain of +5 percentage points over Random Forest	Section 4.3	- Approaches mBERT (87 %); - Outperforms rule-based systems (72–78 %) [31,32,33]
4	Recall_sys +10.4 percentage points Latency increased by only +6.4%	Section 4.3, Section 4.5	Better Recall/Latency gain ratio compared to classical pipelines [1,2,8]
5	Fairness Gap = 4,1 %. USS = 4,5/5	Section 1, Table 2	- Fairness Gap better than without correction (6–12%) [50]; - Higher USS than typical [56]

Hypothesis H3 is also confirmed: the hybrid approach to thematic categorisation (ML ensemble & RAKE/TF-IDF) showed a macro-F1 gain of +5 percentage points compared to Random Forest (Table 4), which meets expectations.

Hypothesis H4, which anticipated an increase in the overall Recall of the final system by at least 10% provided that latency remains within +10%, is supported by the results in Table 5: Recall_sys increased by +10.4 percentage points and latency rose by only +6.4%.

Finally, hypothesis H5 on responsible system evaluation (Fairness Gap ≤ 5%, USS ≥ 4.0) is fulfilled through the computed ΔF1 = 3.7% (Table 5) and USS = 4.5/5 (Table 6). Thus, all points of scientific novelty (1–5) are quantitatively and statistically confirmed.

In interpreting these results, it should be noted that the class-balancing policy (cap ≤ 1,000 per class with down/up-sampling) may alter empirical priors and, in turn, inflate the macro-F1 relative to the field prevalence. A comprehensive prevalence-aware analysis – micro-F1, per-class average precision, calibration to original priors, and class-weighted training without capping – is deferred to future work. The given claims focus on relative module gains and CPU-bound latency/throughput feasibility, which is not expected to be affected by this consideration.

6.2. Comparison of the results with recent works on text mining and Responsible AI

The Responsible-AI (RAIE) layer was explicitly used due the target use case – real-time monitoring of information threats in public-sector and OSINT workflows – carries a non-trivial risk of harm from misclassification and requires auditability, reproducibility, and proportional safeguards. In operational terms, trustworthiness is enforced through deterministic training and inference (fixed seeds, versioned artefacts, audit logs), robustness checks (stratified CV plus out-of-time evaluation and drift monitoring), and fairness auditing (Fairness Gap across languages/sources with thresholds that trigger review). Explainability is provided via Model Cards and local post-hoc attributions, making classification decisions visible to analysts, along with error taxonomies and data statements to contextualise limitations. Low-latency/low-resource constraints are treated as reliability requirements: modules meet CPU-only SLOs (≈45–58 ms/doc per module; pipeline overhead +6–7%) with throughput suitable for streaming, and confidence-thresholding routes low-confidence cases to human-in-the-loop review. The RAIE gates (e.g., ΔF1 ≤ 5%, bounded Fairness Gap, explanation coverage, and latency/throughput SLOs) must be satisfied for deployment; violations surface alerts and block promotion until remediated. This design adheres to the trustworthiness and explainability principles outlined by the co-author in Sensors [70] and aligns with the commonly adopted FATE/AI4People guidelines, while maintaining compatibility with horizontal scaling for peak-load scenarios.

Compared with existing work in text analysis, the proposed system demonstrates competitive and, in some cases, superior results. For example, the sentiment module with macro-F1 = 0.85 outperforms lexicon-based systems such as VADER and SentiWordNet (≈ 0.74–0.76) [21], while the inversion algorithm achieves an MAE improvement of 18.2%, which is greater than the typical 6–12% gains in sarcasm detection systems with negation scope [24]. Thematic classification with a macro-F1 score of 0.83 demonstrates similar effectiveness to fine-tuned mBERT, while retaining a latency of 55 ms, which is 3–4 times lower.

In the area of Responsible AI, the system integrates Model Cards, Fairness Gap, Confidence Thresholding, and User Satisfaction Score, thereby implementing the full FATE principles. Compared to approaches that only implement Post-Processing Equalised Odds [50], the proposed framework reduces the Fairness Gap from 6–12% to 3.7%, with minimal impact on latency. A higher AuditScore (≥ 0.90) was also achieved in model documentation, exceeding the average quality of model cards in open repositories. Therefore, in the context of Responsible AI, the proposed solution not only meets existing standards but also extends their applicability in the field of information security.

These results are consistent with prior research on enterprise-level cybersecurity integration and cognitive decision support systems. Previous studies [68, 69] demonstrated how unified information models and cognitive approaches can enhance decision-making and threat monitoring, laying foundational principles reflected in the current study. Moreover, cognitive modelling facilitates the interpretation of ambiguous or incomplete textual news, which is critical for the early detection of information manipulation. Such approaches are embodied in the context-sensitive sentiment module, the polarity inversion logic and the integration of expert-guided decision thresholds. Together, these components ensure that the system not only achieves high analytical precision but also supports situational awareness in dynamically evolving information spaces.

6.3. Practical implications for information security and politics

The research results have direct practical relevance to building strategic monitoring systems within CERTs, open-source intelligence (OSINT) centres, media monitoring platforms and analytical institutions. Thanks to its

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

integrated architecture (sentiment analysis, inversion and thematic categorisation) and high throughput (up to 17 documents per second) on CPU, the proposed system can be deployed in real-time settings, enabling rapid responses to disinformation campaigns. Such functionality is critical during periods of intensified information attacks or pre-election phases.

In addition to its technical applicability, the system meets both ethical and legal requirements, thereby opening opportunities for its certification under the EU AI Act. Specifically, the implementation of audit trails, documented Model Cards, responsibility metrics (Fairness, Transparency, Accountability), and explainability mechanisms allows the developed solution to be used in politically sensitive environments. The high User Satisfaction Score (4.5/5) further strengthens trust in the system among experts and potential stakeholders, including government agencies and regulatory bodies.

For a more comprehensive quality perspective, future work will benchmark state-of-the-art multilingual transformers on both GPUs and quantised CPUs (INT8) and evaluate distillation pipelines to achieve transformer-level F1 at lexicon-like latency. The macro-F1 with 95% CIs, per-class AP, latency per document, throughput, and resource budgets will be reported to enable fair comparisons.

6.4. Limitations of the study and directions for future work

Despite the successful implementation of the system, the study has certain limitations. The most notable limitation is the corpus: the majority of texts are in Ukrainian, which may limit generalizability to messages in other languages, whereas social media and forums are less represented. In thematic classification, some categories have F1 < 0.80, indicating the need for dictionary refinement and model retraining on specific subgenres.

While appropriate for comparability, stratified 5-fold CV does not fully capture temporal drift or source shift in a 2024 stream. The follow-up experiments with out-of-time splits (e.g., temporal hold-out and rolling-origin schedules) and out-of-source evaluation on held-out media channels to quantify robustness under realistic deployment conditions are then planned.

Future work directions include:

- Multimodal processing (text + image), particularly for memes or pictures in Telegram channels;
- Self-training on large streams of unannotated data to adapt to new topics;
- Improvement of explainability modules through integration of SHAP, counterfactuals and expansion of the Confidence Filtering block with flexible confidence logic. Validation on real-world incidents in partnership with governmental and media institutions is also planned.

The inversion module remains rule-bound and can be brittle on abbreviated anchors and mixed-language inputs, yielding the precision/recall trade-offs reported above. Future work will (i) expand anchor normalisation and source verification and (ii) add a small, learned inversion component to reduce residual false decisions without breaching the end-to-end latency constraints. Alterations to the main conclusions on relative gains and feasibility are not expected.

6.5. Cybersecurity Implications

The developed integrated text mining system significantly advances cybersecurity capabilities by providing real-time detection of hostile information campaigns and disinformation narratives. By enhancing recall and minimising false negatives through polarity inversion detection, the system effectively supports cybersecurity analysts and OSINT experts in identifying coordinated influence operations and early-stage cyber threats. Additionally, the Responsible AI Evaluation ensures compliance with GDPR, transparency, fairness and accountability, thus making the solution highly suitable for deployment in sensitive cybersecurity and governmental contexts. Specifically, the system supports the real-time identification of coordinated hostile narratives, the early detection of information operations preceding cyberattacks, and enables a rapid response during hybrid warfare campaigns.

The system's output can be aligned with MITRE ATT&CK tactics and techniques to contextualise detected narratives within adversarial behaviour models. For instance, specific disinformation themes identified by the thematic classifier correspond to TTPs such as TA0043 (Reconnaissance) or TA0011 (Command and Control) in influence operation contexts. Future work also includes formal mapping to STIX 2.1 objects to enable integration with CTI platforms.

7. Conclusions

In conclusion, this research presents a comprehensive solution that combines high technical efficiency in textual data processing with ethical responsibility in AI, tailored explicitly for cybersecurity tasks such as information threat intelligence, hybrid warfare analytics and CERT operations.

Empirical results confirmed all five hypotheses formulated at the beginning of the study. The context-sensitive sentiment analysis module outperformed baseline models by 7 percentage points in macro-F1 (H1), polarity inversion reduced the mean absolute error by 18.2% (H2), the hybrid thematic classification delivered a +5 percentage point gain over Random Forest (H3), integration of the three modules increased Recall_sys by 10.4 percentage points with acceptable latency growth (H4),

and the comprehensive Responsible AI framework achieved a Fairness Gap of 4.1% and a USS of 4.5/5 (H5), demonstrating full alignment with the stated objectives.

The system's key components include methods for context-sensitive sentiment analysis, polarity inversion detection, thematic classification and Responsible AI metrics (Fairness, Transparency, Accountability, Ethics). The novelty lies in the hybridisation of linguistic and machine learning models, as well as in the integration of quantitative metrics with formalised documentation (Model Cards), ensuring a balance between accuracy and accountability in critical information scenarios.

The system can be deployed in media monitoring analytical centres, cybersecurity structures (CERT, OSINT) and analytical units of government institutions. With a throughput of over 17 documents per second and a latency below 60 ms, it can be scaled for streaming analysis. The presence of an audit trail, Model Cards, and a Confidence Thresholding mechanism enables its integration into ecosystems that comply with the EU AI Act or ISO/IEC 42001 requirements.

While the current implementation focuses on Ukrainian- and Russian-language content, the platform's modular architecture allows for future adaptation to other languages and contexts, provided that appropriate datasets and model retraining procedures are available.

Further research is expected to expand the corpus to include multilingual and multimodal sources, automate model retraining using self-training, develop neural explainability modules (e.g., with SHAP or counterfactual generation), and investigate the dynamics of changes in information narratives over time using drift detection and streaming adaptation models.

Contribution of authors: Hennadii Bohuta, **Lesia Bilovus** and **Khrystyna Yurkiv** compiled a multilingual corpus of news and social-media posts on information threats related to Ukraine for 2022–2025. **Khrystyna Lipianina-Honcharenko**, **Hennadii Bohuta**, **Ihor Ihnatiev** implemented a sentiment analysis module with polarity inversion to detect covert hostility and sarcasm. **Khrystyna Lipianina-Honcharenko**, **Ihor Ihnatiev** developed a hybrid topic classification approach combining dictionary-based methods and machine-learning ensembles. **Oleg Illiashenko**, **Myroslav Komar**, **Ihor Ihnatiev** constructed the Responsible AI Evaluation (RAIE) framework with indicators for Fairness Gap, Model Cards, and User Satisfaction. **Ihor Ihnatiev** integrated all modules into a unified pipeline and performed quality (macro-F1, MAE) and performance (latency, throughput) evaluation. **Khrystyna Lipianina-Honcharenko, Ihor Ihnatiev** conducted the experimental validation of hypotheses H1–H5 and synthesised the resulting evidence. **Oleg Illiashenko** provided the general review and editing of the research results and the manuscript itself.

Project information: This study depicts the results of an interdisciplinary research project between West Ukrainian National University (Department of Information Computer System and Control, and Cyber Security Department), Leeds Beckett University (School of Built Environment, Engineering and Computing), UK, and National Aerospace University "Kharkiv Aviation Institute", Ukraine, dedicated to the development, testing, and implementation of an integrated text-mining technology for real-time monitoring of information threats in Ukraine. Its goal is to design and experimentally validate an end-to-end system that automates context-sensitive sentiment analysis, polarity-inversion detection and thematic classification within a Responsible AI framework. The corpus comprises 10,350 unique documents collected from major Ukrainian news RSS feeds, social media APIs, and relevant blogs during January-December 2024, with stratified 5-fold cross-validation for evaluation. The system is realised as a modular client–server stack (Python + Streamlit + Transformers + Docker) suitable for secure, scalable deployment and integration into compliance-oriented environments (e.g., EU AI Act/ISO/IEC 42001). The authors note that the study was conducted without external financial support.

The technology's effectiveness and reliability hinge on accurate detection of covert manipulations (polarity inversion), fair/transparent classification across multilingual, multi-topic streams and low-latency throughput for operational use (CERT, OSINT, government analytics). Empirically, the sentiment module attains macro-F1 = 0.85; the inversion algorithm reduces polarity MAE by 18.2% with minimal latency overhead; and the hybrid thematic classifier achieves macro-F1 = 0.83 at ≈55 ms/doc and ≈18 docs/s. Integrated end-to-end, the pipeline raises overall recall by +10.4 percentage points while keeping latency growth within ~10% and the RAIE framework ensures $\Delta F1 \leq 5\%$ with an expert User Satisfaction Score of 4.14/5.

The results presented in the paper contribute to addressing critical challenges in information security by providing novel algorithms for context-sensitive sentiment analysis, inversion-aware classification, and responsible evaluation through fairness, accountability, transparency, and user satisfaction indicators.

Conflict of Interest

The authors declare that they have no conflict of interest about this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Financing

This study was conducted without financial support.

Data Availability

The work is accompanied by associated data in the data repository:

https://figshare.com/articles/dataset/News_dataset_about_Ukraine/29020670?file=54415925

Use of Artificial Intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, as described in the research methodology section.

All authors have read and approved the published version of the manuscript.

References

1. Robertson, S., & Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009, vol. 3, iss. 4, pp. 333–389. DOI: 10.1561/1500000019.

2. Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P., Ahmed, J., & Overwijk, A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Proceedings of the International Conference on Learning Representations*, ICLR, 2020. Available at: https://arxiv.org/abs/2007.00808 (accessed 12.05.2025).

3. Stanovsky, G., Michael, J., Zettlemoyer, L., & Dagan, I. Supervised open information extraction. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, USA, ACL, 2018, pp. 885–895. DOI: 10.18653/v1/N18-1081.

4. Hamborg, F., Donnay, K. & Gipp, B. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 2019, vol. 20, pp. 391–415. DOI: 10.1007/s00799-018-0261-y.

5. Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T.H., Ding, K., Karami, M., & Liu, H. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020, vol. 10, iss. 6, article no. e1385. DOI: 10.1002/widm.1385.

6. Doddapaneni, S., Khan, M. S. U. R., Venkatesh, D., Dabre, R., Kunchukuttan, A., & Khapra, M. M. Cross-lingual auto evaluation for assessing multilingual LLMs. *Proceedings of the 64rd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers)*, Vienna, Austria, ACL, 2025, pp. 29297–29329. DOI: 10.18653/v1/2025.acl-long.1419.

7. Pires, T., Schlinger, E., & Garrette, D. How multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, ACL, 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493.

8. Chen, D., Li, Z., Gu, B., & Chen, Z. Multimodal named entity recognition with image attributes and image knowledge. *Database Systems for Advanced Applications*. DASFAA, 2021, *Lecture Notes in Computer Science*, vol. 12682, pp. 183–198. DOI:10.1007/978-3-030-73197-7_12.

9. Satvat, K., Gjomemo, R., & Venkatakrishnan, V.N. Extractor: Extracting attack behavior from threat reports. *Proceedings of the 2021 IEEE European Symposium on Security and Privacy*, EuroS&P, Vienna, Austria, IEEE, 2021, pp. 414–429. DOI:10.1109/EuroSP51992.2021.00046.

10. Kapan, S., & Sora Gunal, E. Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features. *Applied Sciences*, 2023, vol. 13, iss. 24, pp. 1–13, DOI: 10.3390/app132413269.

11. Kondratyuk, D., & Straka, M. 75 languages, 1 model: Parsing Universal Dependencies universally. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* EMNLP-IJCNLP, Hong Kong, China, ACL, 2019, pp. 2779–2795. DOI: 10.18653/v1/D19-1284.

12. Lu, Y., Nie, Z., Cheng, T., Gao, Y., & Wen, J. R. Name disambiguation using a web connection. *Proceedings of AAAI 2007 Workshop on Information Integration on the Web*, IIWeb, Vancouver, Canada, AAAI, 2007, pp. 57–61. Available at: https://cdn.aaai.org/Workshops/2007/WS-07-14/WS07-14-010.pdf (accessed 12.05.2025).

13. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, AAAI, Portland, Oregon, USA, 1996, pp. 226–231. DOI: 10.5555/3001460.3001507.

14. Reimers, N., & Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China, ACL, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

15. Gao, T., Yao, X., & Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, EMNLP, 2021, pp. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.

16. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. A stylometric inquiry into hyperpartisan and fake news. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, ACL, 2018, pp. 231–240. DOI: 10.18653/v1/P18-1022.

17. Corley, C., & Mihalcea, R. Measuring the semantic similarity of texts. *Proceedings of the ACL*

Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, USA, ACL, 2005, pp. 13–18. DOI: 10.3115/1641356.1641359.

18. Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, TODS, 2017, vol. 42, iss. 3, pp. 1–21. DOI: 10.1145/3068335.

19. Gu, X., Angelov, P. P., Kangin, D., & Principe, J. C. A new type of distance metric and its use for clustering. *Evolving Systems*, 2017, vol. 8, iss. 3, pp. 167–177. DOI: 10.1007/s12530-017-9195-7.

20. Artetxe, M., & Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 2019, vol. 7, pp. 597–610. DOI: 10.1162/tacl_a_00288.

21. Hutto, C. J., & Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media*, ICWSM, Ann Arbor, Michigan, USA, AAAI, 2014, vol. 8, no. 1, pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550.

22. Medhat, W., Hassan, A., & Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 2014, vol. 5, no. 4, pp. 1093–1113. DOI: 10.1016/j.asej.2014.04.011.

23. Sun, C., Huang, L., & Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, Minneapolis, Minnesota, USA, Association for Computational Linguistics, 2019, pp. 380–385. DOI: 10.18653/v1/N19-1035.

24. Joshi, A., Bhattacharyya, P., & Carman, M.J. Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 2017, vol. 50, iss. 5, pp. 1–22. DOI: 10.1145/3124420.

25. Kaushik, D., Hovy, E., & Lipton, Z.C. Learning the difference that makes a difference with counterfactually-augmented data. *Proceedings of the 8th International Conference on Learning Representations*, ICLR, 2020. Available at: https://arxiv.org/abs/1909.12434 (accessed 12.09.2025).

26. Volkova, S., & Jang, J.Y. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, Beijing, PRC, ACM, 2018, pp. 575–583. DOI: 10.1145/3184558.3188728.

27. Zhang, L., Wang, S., & Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, vol. 8, no. 4, e1253. DOI: 10.1002/widm.1253.

28. Ghosh, D., & Veale, T. Fracking sarcasm using neural network. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA, San Diego, USA, ACL, 2016, pp. 161–169. DOI: 10.18653/v1/W16-0425.

29. Cakebread‐ Andrews, O., Ha, L. A., Frommholz, I., & Can, B. Error analysis of NLP models and non‐ native speakers of English identifying sarcasm in Reddit comments. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy, ELRA and ICCL, 2024, pp. 6247–6256. Available at: https://aclanthology.org/2024.lrec-main.552/ (accessed 12.05.2025).

30. Breve, B., Caruccio, L., Cirillo, S., Deufemia, V., & Polese, G. Analyzing the worldwide perception of the Russia–Ukraine conflict through Twitter. *Journal of Big Data*, 2024, vol. 11, article no. 76, pp. 1–33. DOI: 10.1186/s40537‐ 024‐ 00921‐ w.

31. Manning, C. D., Raghavan, P., & Schütze, H. *Introduction to Information Retrieval*. Cambridge, Cambridge University Press, 2008. 506 p.

32. Joachims, T. Text categorization with support vector machines: learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, Chemnitz, Germany, AAAI, 1998, pp. 137–142. DOI: 10.1007/BFb0026683.

33. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA, ACL, 2019, vol. 1, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

34. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA, ACL, 2016, pp. 1480–1489. DOI: 10.18653/v1/N16-1174.

35. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

36. Baccianella, S., Esuli, A., & Sebastiani, F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, ELRA, 2010, pp. 2200–2204. Available at: http://lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf (accessed 13.09.2025).

37. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, Vancouver, BD, Canada, IEEE 2019. DOI: 10.48550/arXiv.1910.01108.

38. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. TinyBERT: Distilling BERT for natural language understanding. *Findings of the Association for Computational Linguistics*, 2020, pp. 4163–

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

4174. DOI: 10.18653/v1/2020.findings-emnlp.372.

39. Breiman, L. Random forests. *Machine Learning*. 2001, vol. 45, no. 1, pp. 5–32. DOI: 10.1023/A:1010933404324.

40. Prabowo, R., & Thelwall, M. Sentiment analysis: A combined approach. *Journal of Informetrics*, 2009, vol. 3, no. 2, pp. 143–157. DOI: 10.1016/j.joi.2009.01.003.

41. Forman, G. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*. 2003, vol. 3, pp. 1289–1305. Available at: https://www.jmlr.org/papers/volume3/forman03a/forman03a.pdf (accessed 14.09.2025).

42. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., & Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 2020, vol. 509, pp. 257–289. DOI: 10.1016/j.ins.2019.09.013.

43. Mihalcea, R., & Tarau, P. TextRank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, ACL, 2004, pp. 404–411. Available at: https://aclanthology.org/W04-3252 (accessed 14.05.2025).

44. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, vol. 41, no. 6, pp. 391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

45. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003, vol. 3, pp. 993–1022. Available at: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed 14.05.2025).

46. Grootendorst, M. *KeyBERT: Minimal keyword extraction with BERT*. 2020. Available at: http://dx.doi.org/10.5281/zenodo.4461265 (accessed 14.05.2025).

47. Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. SPECTER: Document-level representation learning using citation-informed transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2020, pp. 2270–2282. DOI: 10.18653/v1/2020.acl-main.207.

48. Wan, X., & Xiao, J. Single document keyphrase extraction using neighborhood knowledge. *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, Chicago, USA, AAAI, 2008, pp. 855–860. Available at: https://cdn.aaai.org/AAAI/2008/AAAI08-136.pdf (accessed 12.05.2025).

49. Papagiannopoulou, E., & Tsoumakas, G. A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2020, vol. 10, no. 2, article no. e1339. DOI: 10.1002/widm.1339.

50. Hardt, M., Price, E., & Srebro, N. Equality of opportunity in supervised learning. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, Curran Associates Inc., 2016, pp. 3323–3331. DOI: 10.5555/3157382.

51. Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., Nair, R., & Nushi, B. Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*. 2019, vol. 63, no. 4/5, pp. 6:1–6:13. DOI: 10.1147/JRD.2019.2942288.

52. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. A. *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping*. 2020. DOI: 10.48550/arXiv.2002.06305.

53. Fort, S., Ren, J., & Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*. NeurIPS, Virtual Conference, Curran Associates Inc., 2021, pp. 1–14. Available at: https://proceedings.neurips.cc/paper_files/paper/2021/file/3941c4358616274ac2436eacf67fae05-Paper.pdf (accessed 14.05.2025).

54. Ribeiro, M. T., Singh, S., & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, ACM, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

55. Lundberg, S. M., & Lee, S.-I. *A* unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. NeurIPS, Long Beach, California, USA, 2017, vol. 30. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (accessed 14.05.2025).

56. Wachter, S., Mittelstadt, B., & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. 2017, vol. 31, no. 2, pp. 841–887. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063289 (accessed 14.05.2025).

57. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*. 2017, vol. 19, no. 1, pp. 22–36. DOI: 10.1145/3137597.3137600.

58. Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language*, Florence, Italy, ACL, 2020, pp. 6025–6044. Available at: https://ora.ox.ac.uk/objects/uuid:3864e746-88c8-4f99-b912-52f4b4be289a/files/m25d33782b006cec944b22e5d744ed1b7 (accessed 14.05.2025).

59. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. AI4People – An ethical framework for a good AI society. *Minds and Machines*. 2018, vol. 28, no. 4, pp. 689–707. DOI: 10.1007/s11023-018-9482-5.

60. Sachenko, A., Lendiuk, T., Lipianina‑Honcharenko, K., Dobrowolski, M., Boguta, G., & Bytsyura, L. Method of determining the text sentiment by thematic rubrics. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems*, Lviv, Ukraine, CEUR, 2024, pp. 404–414. DOI: 10.31110/COLINS/2024-3/026.

61. Lipianina-Honcharenko, K., Soia, M., Yurkiv, K., & Ivasechko, A. Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data. *Proceedings of the 5th International Conference on Computational Methods for Information Security*, Zaporizhzhia, Ukraine, CEUR, 2024, pp. 97–109. Available at: https://ceur-ws.org/Vol-3702/paper9.pdf (accessed 14.05.2025).

62. Lipianina-Honcharenko, K., Lendiuk, D., Melnyk, N., Komar, M., & Lendiuk, T. Evaluation of the keyword selection methods effectiveness for the fake news classification. *Proceedings of the 10th International Scientific Conference on Information Technology and Interactions*, Kyiv, Ukraine, CEUR, 2024, pp. 109–122. Available at: https://ceur-ws.org/Vol-3909/Paper_9.pdf. (accessed 12.05.2025).

63. Zhang, L. L., Han, S., Wei, J., Zheng, N., Cao, T., Yang, Y., & Liu, Y. NN‑Meter: Towards accurate latency prediction of deep‑learning model inference on diverse edge devices. *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. MobiSys, Wisconsin, USA, ACM, 2021, pp. 81–93. DOI: 10.1145/3458864.3467882.

64. Reddi, V.J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.J., Coleman, C., Diamos, G., Elibol, M., Hall, D., Hazelwood, K., Hsu, B., Idiculla, N., Kumar, D., Levenberg, J., Tang, H., Warden, P., & et. al. MLPerf inference benchmark. *Proceedings of the ISCA'20: Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture*. ISCA, Valencia, Spain, IEEE, 2020, pp. 446–459. DOI: 10.1109/ISCA45697.2020.00045.

65. Memarian, B., & Doleck, T. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*. 2023, vol. 5, article no. 100152. DOI: 10.1016/j.caeai.2023.100152.

66. Kharchenko, V., Fesenko, H., & Illiashenko, O. Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic and Computer Systems*. 2022, no. 2, pp. 131–144. DOI: 10.32620/reks.2022.2.11

67. *Streamlit. Streamlit web application*. Available at: https://github.com/streamlit/streamlit (accessed 14.05.2025).

68. Dotsenko, S., Illiashenko, O., Kharchenko, V., & Morozova, O. Integrated information model of an enterprise and cybersecurity management system: From data to activity. *International Journal of Cyber Warfare and Terrorism*. 2022, vol. 12, no. 2, pp. 1–21. DOI: 10.4018/IJCWT.305860.

69. Mygal, V., Mygal, G., & Illiashenko, O. Intelligent decision support – cognitive aspects. In: *Digital Transformation, Cyber Security and Resilience of Modern Societies. Studies in Big Data*, vol. 84. Cham, Springer, 2021, pp. 395–411. DOI: 10.1007/978-3-030-79934-2_26.

70. Kharchenko, V., Fesenko, H., & Illiashenko, O. Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. *Sensors*, 2022, vol. 22, iss. 13, article no. 4865. DOI: 10.3390/s22134865.

Received 15.05.2025, Accepted 25.08.2025

ЗАГАЛЬНИЙ МЕТОД ВИЯВЛЕННЯ ІНФОРМАЦІЙНИХ ЗАГРОЗ У РЕЖИМІ РЕАЛЬНОГО ЧАСУ НА ПРИКЛАДІ УКРАЇНИ

Х. В. Ліп'яніна-Гончаренко, М. П. Комар, Г. М. Богута, І. В. Ігнатєв, Х. В. Юрків, О. О. Ілляшенко, Л. І. Біловус

Предметом дослідження є загальний набір методів і системна архітектура для текстової аналітики, що забезпечують виявлення та моніторинг інформаційних загроз у режимі реального часу, верифіковані на кейсі України. Запропонований набір методів інтегрує аналіз тональності, оброблення інверсії полярності та тематичну класифікацію на основі методів машинного навчання. Дослідження актуалізується в умовах гібридної війни, коли інформаційне середовище стає полем дезінформації, маніпулятивних кампаній та когнітивного впливу. **Метою** є розробка та експериментальна валідація комплексної інформаційної технології для автоматизованого виявлення загроз в інформаційному просторі України, яка базується на принципах відповідального штучного інтелекту (Responsible AI) та сучасних методах обробки природної мови. **Завдання**: формування багатомовного корпусу текстів новин та соціальних медіа, реалізація модуля аналізу тону з урахуванням інверсії полярності, розробка гібридного методу тематичної класифікації із залученням словників ключових слів та ансамблів моделей машинного навчання, побудова фреймворку оцінювання відповідального ШІ (RAIE) із показниками чесності, прозорості та користувацької задоволеності. Отримані **результати** підтверджують всі п'ять висунутих гіпотез: розроблений модуль аналізу тону досягає макро-F1 = 0.85 та знижує MAE на 18.2% порівняно з базовою моделлю; алгоритм виявлення інверсії дозволяє автоматично змінювати знак емоційного індексу при виявленні маніпулятивних повідомлень, покращуючи точність визначення ворожих наративів; гібридна тематична класифікація забезпечує макро-F1 = 0.83 при затримці 55 мс/документ та продуктивності 18 документів/секунда; інтеграція модулів у єдиний пайплайн підвищує повноту виявлення загроз на 10.4% без істотного зростання затримки; впровадження концептуальної моделі RAIE забезпечує ΔF1

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

≤ 5%, середній бал задоволеності експертів 4.14/5 та незначне збільшення затримки (<10%). Висновки свідчать про те, що запропонована система поєднує високу точність виявлення інформаційних загроз із принципами етичності, прозорості та користувацької довіри, що забезпечує її практичну цінність для державних центрів кіберзахисту, CERT та OSINT-платформ. **Висновки.** Наукова новизна полягає у розробці нових методів: контекстно-чутливого аналізу тону з урахуванням специфіки військової лексики; алгоритму інверсії полярності для виявлення прихованої ворожості; гібридної тематичної класифікації, що поєднує машинне навчання та експертні словники; інтегрованої архітектури інформаційної технології з продуктивністю >17 документів/секунда; моделі оцінювання відповідального ШІ з впровадженням Fairness Gap, Model Cards та User Satisfaction Score.

Ключові слова: текстовий майнінг; аналіз настроїв; виявлення інверсій; класифікація текстів; машинне навчання.

Ліп'янінa-Гончаренко Христина Володимирівна – д-р техн. наук, доц., Західноукраїнський національний університет, Тернопіль, Україна.

Комар Мирослав Петрович – д-р техн. наук, проф., Західноукраїнський національний університет, Тернопіль, Україна.

Богута Геннадій Миколайович – студент, Західноукраїнський національний університет, Тернопіль, Україна.

Ігнатєв Ігор Васильович – викладач, Західноукраїнський національний університет, Тернопіль, Україна.

Юрків Христина Володимирівна – студентка, Західноукраїнський національний університет, Тернопіль, Україна.

Ілляшенко Олег Олександрович – канд. техн. наук., Університет Лідс Беккет, Лідс, Велика Британія; доц., каф. комп'ютерних систем, мереж та кібербезпеки факультету радіоелектроніки, комп'ютерних систем та інфокомунікацій, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Біловус Леся Іванівна– д-р іст. наук, проф., Західноукраїнський національний університет, Тернопіль, Україна.

Khrystyna Lipianina-Honcharenko – Doctor of Technical Sciences, Associate Professor Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: xrustya.com@gmail.com, ORCID: 0000-0002-2441-6292, Scopus Author ID: 59548850400.

Myroslav Komar – Doctor of Technical Sciences, Professor Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: mko@wunu.edu.ua, ORCID: 0000-0001-6541-0359, Scopus Author ID: 35366491300.

Hennadii Bohuta – Student, Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: genaboguta7@gmail.com, ORCID: 0009-0000-9788-1753, Scopus Author ID: 59155725000.

Ihor Ihnatiev – Lecturer, Department, of Cyber Security, West Ukrainian National University, Ternopil, Ukraine,
e-mail: iiv@wunu.edu.ua, ORCID: 0000-0003-0729-9247, Scopus Author ID: 57191954223

Khrystyna Yurkiv – Student, Department of Information Computer System and Control, West Ukrainian National University, Ternopil, Ukraine,
e-mail: kh.yurkiv@wunu.edu.ua, ORCID:0009-0007-4917-3251, Scopus Author ID: 58776326000.

Oleg Illiashenko – PhD, School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, United Kingdom; Associate Professor, Department of Computer Systems, Networks and Cybersecurity, Faculty of Radio Electronics, Computer Systems and Infocommunications, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine;
e-mail: o.illiashenko@leedsbeckett.ac.uk, o.illiashenko@khai.edu, ORCID: 0000-0002-4672-6400, Scopus Author ID: 55842633400.

Lesia Bilovus – Doctor of History, Professor of the Department of Information and Socio-Cultural Activity, West Ukrainian National University, Ternopil, Ukraine,
e-mail: l.bilovus@wunu.edu.ua, ORCID: 0000-0003-4882-4511, Scopus Author ID: 57219598554.