МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ В УКРАЇНСЬКИХ ТЕКСТОВИХ ДАНИХ

Колективна монографія

Тернопіль ЗУНУ 2025

УДК 004:8

І 73

Розглянуто та рекомендовано до друку на засіданні Вченої ради Західноукраїнського національного університету (протокол № 1 від 07.05.2025 р.).

І 73 Інтелектуальні методи виявлення дезінформації в українських текстових даних / за ред. д. т. н., доцент, Х. В. Ліп'яніна-Гончаренко. Тернопіль: Економічна думка, 2025. 200 с.

ISBN 978-966-654-779-1

Монографія досліджує сучасні методи виявлення дезінформації в українських текстових даних із використанням машинного навчання. Розглянуто актуальні виклики інформаційної безпеки, гібридних загроз та методів моніторингу новинних ресурсів і соціальних мереж. Запропоновані підходи сприятимуть зміцненню довіри до інформаційних джерел і створенню надійного інформаційного середовища. Результати дослідження будуть корисними науковцям, фахівцям із кібербезпеки та аналізу даних.

ISBN 978-966-654-779-1

ЗМІСТ

Передмова 7

1 ФЕЙКИ ЯК ОСНОВНА ЗАГРОЗА УПРАВЛІННЯ БЕЗПЕКОВОЮ КОМПОНЕНТОЮ: АКТУАЛІТЕТ ТА ВЕКТОРИ ПРОТИДІЇ 12

1.1 Роль фейкової інформації у підриві державної безпеки.....12
1.2 Класифікація та характеристики фейкових новин.....15
1.3 Стратегії запобігання трансформаційно-руйнівним процесам у державі.....18
1.4 Висновки з проведеного дослідження.....21
Список використаних джерел: 22

2 ШТУЧНИЙ ІНТЕЛЕКТ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВОЇ ІНФОРМАЦІЇ: МІЖНАРОДНІ ТРЕНДИ І МОЖЛИВОСТІ ІМПЛЕМЕНТАЦІЇ.....24

2.1 Актуальність дослідження та його значущість.....24
2.2 Постановка проблеми та аналіз сучасного стану досліджень .25
2.3 Виклад основного матеріалу дослідження.....29
2.4 Аналіз ефективності методів виявлення фейкової інформації за допомогою машинного навчання.....36
2.5 Висновки з проведеного дослідження.....40
Список використаних джерел: 42

3 ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ В УКРАЇНСЬКИХ ТЕКСТОВИХ ДАНИХ.....45

3.1 Огляд існуючих рішень 45
3.2 Методологія дослідження.....48
3.2.1 Опис датасету 48

3.2.2 Опис використаних класифікаторів 51
3.2.3 Тренування моделей 56
3.3 Результати дослідження 58
3.4 Висновки з проведеного дослідження.....64
Список використаних джерел: 66
4 ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ ВИДІЛЕННЯ КЛЮЧОВИХ СЛІВ ДЛЯ КЛАСИФІКАЦІЇ ФЕЙКОВИХ НОВИН.....69
4.1 Аналітичний огляд та методологія дослідження фейкових новин.....69
4.2 Методологія дослідження.....72
4.2.1 Архітектура дослідження 72
4.2.2 Опис датасетів.....73
4.2.3 Опис використаних виділених ключових слів 75
4.2.4 Класифікатор 84
4.3 Результати дослідження 85
4.4 Висновки з проведеного дослідження.....90
Список використаних джерел: 91
5 МЕТОД ВИЗНАЧЕННЯ НАСТРОЮ ТЕКСТУ ЗА ТЕМАТИЧНИМИ РИБРИКАМИ.....94
5.1 Розробка методу аналізу тональності тексту в контексті інформаційної війни 94
5.2 Метод 97
5.3 Реалізація 99
5.4 Висновки з проведеного дослідження.....104
Список використаних джерел: 105
6 АНСАМБЛЬ КЛАСИФІКАТОРІВ ДЛЯ ОНЛАЙН ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ.....108
6.1 Аналіз попередніх досліджень 108
6.2 Метод ансамблю класифікаторів для онлайн виявлення дезінформації.....119

6.3 Реалізація 124
6.4 Висновки з проведеного дослідження.....128
Список використаних джерел: 129
7 ГІБРИДНИЙ ПІДХІД ДО ВИЗНАЧЕННЯ ДІПФЕЙКІВ НА ОСНОВІ ЛЮДСЬКОГО ОБЛИЧЧЯ.....135
7.1 Сучасні підходи до визначення діпфейків 135
7.2 Фактори, що визначають діпфейк.....137
7.3 Узагальнений підхід до визначення діпфейків 138
7.4 Висновки з проведеного дослідження.....142
Список використаних джерел: 142
8 ПОРІВНЯЛЬНИЙ АНАЛІЗ АРХІТЕКТУР CNN ДЛЯ КЛАСИФІКАЦІЇ ЕМОЦІЙНИХ СТАНІВ НА ЛЮДСЬКОМУ ОБЛИЧЧІ.....144
8.1 Порівняння архітектур ЗНМ для класифікації емоцій за зображеннями обличчя.....144
8.2 Методи класифікації емоцій за зображеннями облич 145
8.3 Узагальнений алгоритм класифікації зображень.....146
8.4 Аналіз архітектур згорткових нейронних мереж.....147
8.5 Комп’ютерні експерименти.....150
8.1 Висновки з проведеного дослідження.....153
Список використаних джерел: 155
9 АНАЛІЗ ЗАСОБІВ ВИЯВЛЕННЯ DEEPFAKE НА ОСНОВІ АНАЛІЗУ ОБЛИЧЧЯ НА ВІДЕО.....157
9.1 Сучасні підходи до визначення діпфейків на відео 157
9.2 Інструменти для пошуку діпфейків 160
9.3 Висновки з проведеного дослідження.....165
Список використаних джерел: 165
10 Нейромережевий ансамблевий метод до класифікації діпфейків із використанням вибору золотих кадрів 168

10.1 Актуальність проблеми та огляд сучасних методів детекції діпфейків.....168
10.2 Розробка та реалізація ансамблевого методу.....176
10.3 Реалізація.....186
10.3.1 Case Study.....186
10.3.2 Вплив ознак.....190
10.3.3 Оцінки точності.....192
10.3.4 Опис модуля та його роботи.....193
10.4 Обговорення.....194
10.5 Висновки з проведеного дослідження.....196
Список використаних джерел:.....197

ПЕРЕДМОВА

Фейкова інформація та дезінформація останніми роками набули значних масштабів і стали серйозною загрозою для сучасного суспільства. Це явище особливо небезпечне в умовах інформаційних воєн, коли фейкові новини стають засобом маніпуляції громадською думкою, поширюються через соціальні мережі та новітні цифрові канали й здатні викликати масову дестабілізацію. Інформаційні атаки можуть мати довготривалі наслідки, оскільки вони впливають на свідомість громадян, формуючи викривлене уявлення про реальність, що підриває соціальну єдність та взаємодовіру в суспільстві.

У контексті України проблема дезінформації набуває особливої гостроти. Постійні спроби зовнішніх сил використовувати інформаційні ресурси для маніпулювання українським суспільством мають на меті підрив довіри до державних інституцій, розпалювання внутрішніх конфліктів та поляризацію населення за мовними, культурними і релігійними ознаками. Такі дії становлять пряму загрозу національній безпеці й можуть спричиняти значні соціальні, економічні та політичні наслідки, зокрема, шляхом поглиблення регіональних суперечок і зниження ефективності державного управління.

З огляду на це, розвиток методів і технологій для автоматичного виявлення фейкової інформації набуває критичного значення. Використання машинного навчання в цій сфері дозволяє розробляти системи для швидкої ідентифікації дезінформації, що сприятиме зміцненню інформаційної безпеки України. Створення таких технологій дозволить знизити вплив шкідливих інформаційних потоків на громадську думку, підвищити довіру до офіційних джерел і забезпечити формування стійкого, захищеного інформаційного середовища.

Метою даної монографії є розробка та аналіз теоретичних і методологічних підходів до виявлення фейкової інформації за допомогою методів машинного навчання та штучного інтелекту, а також надання практичних рекомендацій для їх застосування в українському контексті. Основна увага приділяється створенню інструментів, які допоможуть у боротьбі з дезінформацією та покращать здатність держави й суспільства протидіяти негативному впливу фейкових новин.

Для вирішення поставленої мети, потрібно виконати наступні ***завдання***:

1. Провести огляд наукових джерел та визначити актуальні проблеми, пов'язані з поширенням фейкової інформації в Україні та світі.
2. Розробити класифікацію фейкової інформації та визначити її основні характеристики для побудови ефективних алгоритмів виявлення.
3. Дослідити методи штучного інтелекту та машинного навчання, що застосовуються для ідентифікації дезінформації, та оцінити їх ефективність.
4. Розробити та апробувати методи виявлення фейкових новин на основі аналізу тексту та виділенні ключових слів.
5. Розробити методику визначення настрою тексту та застосувати її для аналізу фейкових новин у контексті зміцнення інформаційної безпеки.
6. Оцінити ефективність ансамблевих класифікаторів для онлайн-виявлення дезінформації та протестувати їх у реальних умовах.
7. Запропонувати практичні рекомендації для використання розроблених методів у різних галузях, зокрема для забезпечення інформаційної безпеки держави та протидії гібридним інформаційним загрозам.

Наукова новизна дослідження полягає у розробці інноваційних методів для автоматизованого виявлення фейкової інформації, адаптованих до специфіки українських текстових даних. На відміну від існуючих рішень, запропонований підхід враховує не лише текстові характеристики, але й емоційні та семантичні ознаки, що дозволяє підвищити точність і надійність виявлення дезінформації. Запропоновано нові методи побудови ансамблю класифікаторів для онлайн-виявлення фейкової інформації, що дозволяє відстежувати та аналізувати динаміку її поширення в реальному часі. Дослідження також включає авторську класифікацію фейкової інформації за ступенем її впливу на різні соціальні групи, що сприяє кращому розумінню ефектів дезінформації на суспільну думку та розробці ефективніших методів протидії.

Результати монографії мають широке ***практичне*** застосування в таких сферах, як інформаційна безпека, державне управління, медіа-аналітика та соціологія. Зокрема, розроблені методи можуть бути використані державними установами для виявлення дезінформації та протидії негативним інформаційним впливам на суспільну думку. Інструменти для автоматичного виявлення фейкової інформації можуть бути інтегровані в системи моніторингу соціальних медіа та новинних ресурсів для своєчасного реагування на загрози, що виникають у медіапросторі. У сфері освіти й медіаграмотності результати можуть бути застосовані для підвищення рівня критичного мислення громадян щодо споживання інформації. Крім того, практичне значення розроблених методів включає можливість їх адаптації до інших мов та культурних контекстів, що робить результати монографії корисними для міжнародної спільноти в питаннях інформаційної безпеки.

Монографія складається з дев’яти розділів, які логічно пов’язані для досягнення поставленої мети дослідження. Перший розділ, «Фейки як

основна загроза управління безпековою компонентою: актуалітет та вектори протидії», описує загальні аспекти дезінформації, її вплив на державну безпеку та основні напрями протидії. У другому розділі, «Штучний інтелект для виявлення фейкової інформації: міжнародні тренди і можливості імплементації», аналізуються сучасні методи виявлення фейкової інформації на основі штучного інтелекту та оцінюється їх актуальність для українського контексту. Третій розділ, «Оцінка ефективності методів машинного навчання для виявлення дезінформації в українських текстових даних», присвячений дослідженню та порівнянню різних моделей машинного навчання для класифікації фейкових новин. Четвертий розділ, «Оцінка ефективності методів виділення ключових слів для класифікації фейкових новин», аналізує методи виділення ключових слів та їх вплив на якість класифікації. П'ятий розділ, «Метод визначення настрою тексту за тематичними рубриками», описує розроблений підхід для аналізу тональності тексту, який можна використовувати для виявлення дезінформаційних матеріалів. Шостий розділ, «Ансамбль класифікаторів для онлайн виявлення дезінформації», де пропонується модель ансамблевих класифікаторів для швидкого та ефективного виявлення фейкових новин у реальному часі. Сьомий розділ, «Гібридний підхід до визначення діпфейків на основі людського обличчя», охоплює сучасні підходи до виявлення діпфейків за аналізом обличчя, зокрема, різні техніки та фактори, що впливають на точність ідентифікації. У восьмому розділі, «Аналіз засобів виявлення діпфейків на основі аналізу обличчя на відео», розглядаються інструменти та алгоритми, призначені для виявлення діпфейків у відеоконтенті, оцінюється їх ефективність і надійність. Завершує роботу дев’ятий розділ, «Порівняльний аналіз архітектур CNN для класифікації емоційних станів на людському обличчі», де проводиться порівняння різних архітектур згорткових нейронних мереж для

класифікації емоцій за зображеннями обличчя, що забезпечує всебічне розуміння точності, швидкодії та застосувань кожної моделі.

Кожен розділ доповнює попередній, забезпечуючи комплексне розкриття теми дезінформації та реалізації інтелектуальних методів для її виявлення, що сприяє досягненню мети дослідження.

1

ФЕЙКИ ЯК ОСНОВНА ЗАГРОЗА УПРАВЛІННЯ БЕЗПЕКОВОЮ КОМПОНЕНТОЮ: АКТУАЛІТЕТ ТА ВЕКТОРИ ПРОТИДІЇ

Вівчар Оксана[1], Дракохруст Тетяна[1], Тюхтенко Наталія[2]
[1]Західноукраїнський національний університет, вул. Львівська, 11,
м. Тернопіль, 46000
[2]Херсонський державний університет, вул. Університетська, 27,
м. Херсон, 73003
Україна

1.1 Роль фейкової інформації у підриві державної безпеки

Зазвичай, утворення держави відбувається за волею народу і супроводжується різними негативними дезінформаційними явищами, серед яких поширення фейкових новин. Успішний розвиток держави значною мірою залежить як від професійної підготовки її керівників та здатності протидіяти дезінформації, так і від рівня свідомості громадян, їхніх професійних якостей, патріотизму та порядності. Тому деякі держави, незважаючи на невеликі розміри, здобувають авторитет, силу та вплив у міжнародній спільноті. Інші держави, попри великі масштаби, через вищезгадані фактори стикаються з численними протиріччями і розвиваються менш ефективно. Отже, в умовах сучасних загроз фейкова інформація не лише не зміцнює відчуття безпеки серед населення, а й створює загрозу, яка негативно впливає на військово-політичну та економічну стабільність в Україні. Поширення фейкових новин може бути зумовлене як навмисною дезінформацією з боку противника, так і

інформаційною необізнаністю, слабкою виховною роботою та відсутністю чіткої державної ідеології.

Проведені дослідження показали, що термін "фейк" має іншомовне походження (з англійської «fake»), що означає підробку або фальсифікацію. Спочатку цей термін з’явився в інтернет-середовищі, а згодом набув широкого поширення у повсякденному вжитку. Фейк являє собою навмисно створену новину, подію або журналістський матеріал, що містить неправдиву або перекручену інформацію, яка може дискримінувати як окрему людину чи групу осіб в очах громадськості, так і все суспільство. Фейки можуть відрізнятися за формою, методами поширення та змістом. За способом поширення розрізняють: масмедійні фейки (спеціально створені для розповсюдження через ЗМІ) та мережеві чутки (вигадки, які поширюються через соціальні мережі). За формою виділяють: фотофейки, відеофейки та фейкові журналістські матеріали.

Проблема пошуку шляхів підвищення державної безпеки та запобігання поширенню фейкової інформації, пов'язаної з можливими антидержавними змовами, негативно впливає на ефективність виробництва, фінансову стійкість, психологічну стабільність громадян України та діяльність господарських суб’єктів. Слід зазначити, що ця тематика частково висвітлена в наукових працях таких дослідників, як О. Барановський, І. Бінько, Т. Васильців, О. Вівчар, В. Геєць, М. Єрмошенко, В. Мунтіян, І. Назаренко, І. Рекун, В. Сичов, А. Ткач, А. Шастітко, В. Шлемко, В. Якубенко та інших.

Метою цієї наукової роботи є розробка теоретико-методичних підходів і практичних рекомендацій для запобігання поширенню фейкової інформації про антидержавні змови, що сприятиме зміцненню системи національної безпеки.

Не можна залишити без уваги той факт, що байдужість, некомпетентність і злочинність стають інструментами досягнення амбітних цілей як регіональних, так і глобальних лідерів, а також малоосвічених громадян України. Внаслідок цього зникали видатні особистості, талановиті люди, цілі народи, і це супроводжувалося стражданнями та муками. Згодом, коли суспільство заспокоювалося після внутрішніх конфліктів, досягнення та винахідливість завойовників, що нерідко базувалися на підступності, починали прославлятися амбіційними особистостями. Нове покоління, часто схильне до спотвореного бачення історії та політики, нав'язувало суспільству образ геніальності й таланту завойовників, навіть якщо ці досягнення ґрунтувалися на жорстокості та маніпуляціях.

Варто зазначити, що агресор, використовуючи гібридні методи, такі як фейки, плітки та чутки, намагається поляризувати українське суспільство на мовному ґрунті та спровокувати міжнаціональні конфлікти. Наприклад, у деяких регіонах спроби спрямовані на розпалювання протиріч між українцями та угорцями (венграми), в інших — з поляками, а також намагання викликати конфлікт із кримськими татарами, які, втім, виявили більшу мудрість, ніж очікували агресори. У центральних регіонах України зусилля спрямовані на розпалювання ворожнечі проти євреїв, які нині займають помітні позиції в керівництві держави. Проте, варто зазначити, що євреї завжди були частиною українського суспільства, як і багатьох інших протягом історії людства. За їхньої участі розквітали й занепадали держави, і вони переживали як знищення, так і відродження. Тисячолітні пророцтва справджувалися, але історія продовжувала свій рух, часто не залишаючи уроків. Це реальність нашого сьогодення. Шизофренічні теорії змови, нав’язані українцям, грають на руку агресору, який і без того відкрито діє проти них.

1.2 Класифікація та характеристики фейкових новин

Як відомо, фейки з дезінформаційним характером можна поділити на такі категорії: абсолютно недостовірні; недостовірні з елементами правдоподібності; достовірні чутки з елементами неправдоподібності; та правдоподібні. Фейки ніколи не бувають повністю достовірними, адже інформація завжди має певну мету, спрямовану на конкретну аудиторію і змінюється залежно від того, хто її подає та поширює. За силою впливу фейки поділяються на ті, що орієнтовані на формування групової думки, на ті, що провокують індивідуальні або масові антисоціальні настрої, та на ті, що руйнують зв'язки між особами та групами.

Згідно з цією класифікацією, можна простежити, як саме дезінформація впливає на стосунки між індивідами. Наприклад, якщо фейкова інформація має негативний характер, вона може руйнувати зв’язки всередині групи. Водночас чутка, яка не несе емоційного заряду, не матиме впливу на групу.

За походженням фейки поділяються на стихійні [3], тобто ті, що виникають спонтанно без чіткої причини, та навмисно сфабриковані, створені для досягнення конкретної мети. Як правило, навмисно створені чутки поширюються швидше у групі та можуть мати на неї негативний вплив. За масштабом поширення фейкова інформація може охоплювати як невелику групу людей, так і широку аудиторію.

Процес передачі інформації та її поширення демонструє, що для цього достатньо хоча б двох осіб, які взаємодіють між собою, передаючи так звану конфіденційну інформацію. Це можна пояснити так: той, хто отримує інформацію, відповідає тому, хто її надає, і в процесі цієї взаємодії вони спільно створюють новий зміст.

Також можна виділити два типи взаємодії [6]: короткочасну та довгострокову. Короткочасна взаємодія швидко забувається і майже не має

наслідків. Натомість, довгострокова взаємодія може мати такі характеристики, як антагоністичність або солідарність. Антагоністична взаємодія виникає, коли одна сторона діє протилежно іншій. Наприклад, коли людина (А) поширює чутку про людину (Б), і в свою чергу людина (Б) заперечує її правдивість, відповідаючи новою чуткою про людину (А).

Фейкову інформацію також класифікують за інтенсивністю її поширення на продуктивну та непродуктивну. Наукові джерела вказують, що «інтенсивність поширення фейків прямо пропорційна інтересу аудиторії до теми та обернено пропорційна кількості офіційних повідомлень на цю тему та рівню довіри до джерела інформації. Це твердження стосується лише умов, за яких виникають і поширюються чутки. З точки зору функціональних мотивів, циркуляція чуток сприяє міжособистісним контактам, надаючи їм додаткового імпульсу» [2].

Фейки, які містять хибну інформацію, мають високу ймовірність виникнення. Встановлено, що в умовах гібридної війни та інших агресивних дій [1], дезінформація та фейки під час поширення можуть зазнавати різних трансформацій.(Рисунок 1.1)

Наявність такої дезінформації дозволяє виявити ключові ознаки розвитку трансформаційно-руйнівних процесів у державі, розробити концепцію організаційного механізму їх виникнення, а також запобігти негативним наслідкам цих трансформаційно-руйнівних (виробничих) та деградаційних (свідомісних) проявів.

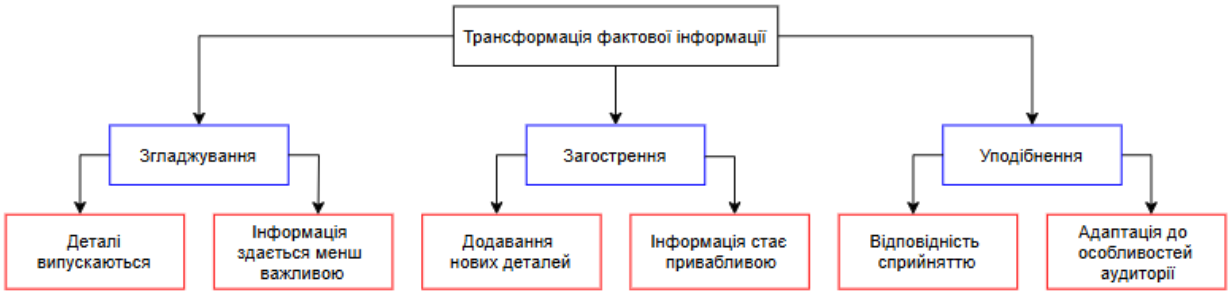


Рисунок 1.1. Процес трансформації фейкової інформації

Без особливих зусиль можна виявити ознаки або передумови розвитку трансформаційно-руйнівних і деградаційних процесів у державі [4] (Рисунок 1.2), які часто викликані поширенням фейкової інформації, пліток та чуток.

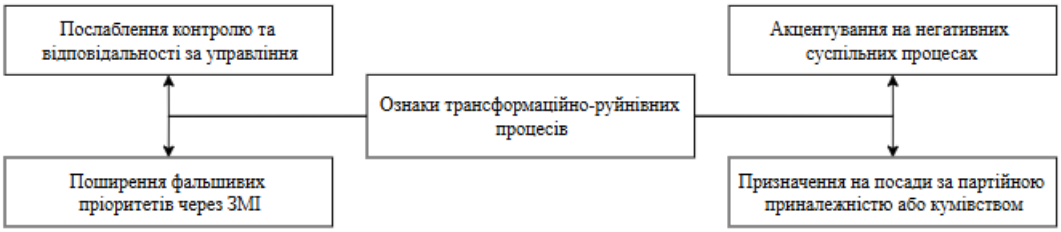


Рисунок 1.2. Ознаки трансформаційно-руйнівних процесів

Варто зазначити, що характерними ознаками проявів трансформаційно-руйнівних і деградаційних процесів у державі є, зокрема:

- спотворення основних функцій держави таких, як управлінська, виховна, юридично-правова, соціально-захисна, оборонна та інші, які реалізуються через певні механізми та засоби;
- нівелювання ключових функцій держави та визнання її відповідальним органом за виховання населення;
- виправдання негативних явищ шляхом порівняння з ще більшими проблемами за кордоном; провокування безглуздих процесів за прикладом сусідніх країн.

1.3 Стратегії запобігання трансформаційно-руйнівним процесам у державі

Агресор намагається всіма засобами дестабілізувати суспільну ситуацію, сподіваючись, що спровоковані конфлікти призведуть до громадянської війни. На окупованих територіях ворог завдає значної шкоди виробничій сфері, незаконно вивозячи ресурси та знищуючи інфраструктуру, таку як дороги, мости і житло. Своїми фейковими втручаннями в Інтернет агресор забруднює інформаційний простір не лише в межах держави, а й на глобальному рівні. Ворог також здійснюють диверсії у фінансовій та банківсько-кредитній системах, продаючи свої енергоресурси одній стороні за заниженими цінами для здобуття прихильності, а іншій — за завищеними, що посилює непорозуміння і викликає ненависть до їхніх дій.

Шляхи запобігання трансформаційно-руйнівним і деградаційним процесам у державі, усунення фейкової інформації та активізація виходу суспільства з кризи повинні включати вдосконалення системи державного управління, а також пошук національної ідеї, яку готові беззастережно підтримувати не лише політики та державні службовці, а й більшість населення. Це також передбачає розробку прогнозів та планів розвитку національної економіки як на короткострокову перспективу, так і на майбутнє. Виконавчі органи мають запропонувати організаційно-економічні механізми для реалізації та впровадження цих планів[5]. Така комплексна мета вимагає створення необхідної бази, що передбачає виконання наступних завдань (Рисунок 1.3).

Рисунок 1.3. Шляхи запобігання трансформаційно-руйнівним процесам та активізація виходу з кризи

Виявлення негативних ознак процесів (з урахуванням чинників), які призводять до трансформаційно-руйнівних і деградаційних явищ, дає змогу відстежувати динаміку цих процесів та вживати необхідних управлінських заходів для запобігання їх негативним наслідкам.

Як показують соціологічні дослідження, змова виникає через переоцінку людиною власних можливостей. Її реалізація призводить до значних жертв і супроводжується великим духовним покаранням. Дійсно, немає нічого прихованого, що не стане відомим.

Таким чином, вирішення проблеми поширення дезінформації та фейків як одного з аспектів забезпечення національної безпеки в умовах гібридної війни сприятиме не лише запобіганню їх негативним наслідкам, але й забезпеченню економічної стабільності та спокою в суспільстві. Соціологічні узагальнення та рекомендації розроблено в контексті активізації державотворення відповідно до її функцій.

Виконання державою своїх функцій у цій сфері має постійний та систематичний характер і триває протягом усього періоду існування, відповідно до завдань, що стоять перед нашою незалежною Україною. Реалізація цих функцій залежить від низки чинників, серед яких важливу роль відіграє визначення напрямку розвитку держави, що здійснюється безпосередньо її керівництвом. Наприклад, наразі свідомо ігнорується та не

виконується виховна функція держави, яка могла б значною мірою зменшити негативний вплив фейків та пліток на суспільство. Відсутність формування державної ідеології значно впливає на розвиток патріотичної свідомості у молоді, залишаючи населення без оптимізму та перспективи. Дослідники вказують, що в країнах, де відсутня ідеологічна робота, рівень самогубств на 30% вищий порівняно з тими країнами, де люди мають надію на краще майбутнє.

Основні функції держави за сферами діяльності поділяються на внутрішні та зовнішні. Внутрішні функції виконуються всередині країни та відображають її внутрішню політику. Вони охоплюють політичну, економічну, екологічну сфери, забезпечення та захист прав і свобод людини, культурну, виховну, освітню, наукову, спортивну та пропагандистську діяльність, соціальну сферу, підтримку громадського порядку тощо. Зовнішні функції держави спрямовані на реалізацію її зовнішньої політики.

Основними формами здійснення державних функцій є правові та фактичні (організаційні) форми. До головних правових форм відносяться: правотворча, правозастосовна, правоохоронна, контрольно-наглядова, інтерпретаційно-правова, управлінська та засновницька діяльність. Важливо зазначити, що для детального та всебічного аналізу виконання будь-якої функції держави, особливо виховної, необхідно розглядати її у контексті змісту, форм та методів реалізації відповідного напряму державної діяльності.

Для зменшення та запобігання негативним наслідкам дезінформації та фейкових новин, зумовлених невизначеністю повідомлень, а також для підвищення якості інформаційного забезпечення, доцільно розробити модель ефективного використання сучасних інформаційних технологій. Ця модель повинна включати розвинену систему управління інформацією,

методи оцінки фактичного та оптимального фінансово-економічного стану за допомогою інтегральних показників, а також нові інструменти для моніторингу та діагностики економічної безпеки стратегічних підприємств країни.

1.4 Висновки з проведеного дослідження

У дослідженні підтверджено, що фейкова інформація є однією з найнебезпечніших загроз для системи державної безпеки, особливо в умовах гібридної війни. Розповсюдження фейків дестабілізує суспільство, підриває довіру до інституцій, провокує міжетнічні конфлікти та перешкоджає ефективному функціонуванню органів влади. Фейки, що мають різні форми (текстові, візуальні, відео), активно використовуються як інструмент інформаційної агресії проти України.

Проведений аналіз дозволив класифікувати фейкову інформацію за рівнем достовірності, джерелом виникнення, масштабом впливу та інтенсивністю поширення. Визначено, що найнебезпечнішими є навмисно сфабриковані фейки, орієнтовані на руйнування соціальних зв'язків і провокування деструктивних настроїв у суспільстві. Зафіксовано, що такі фейки швидко трансформуються, змінюючи форму та зміст, що ускладнює процес їх ідентифікації та протидії.

Окреслено ознаки трансформаційно-руйнівних процесів, що виникають внаслідок фейкової агресії, серед яких – спотворення ключових функцій держави, підрив довіри до правових та виховних механізмів, та зниження рівня патріотизму. Особливо загрозливою є відсутність цілісної державної ідеології, яка б могла стати основою для протистояння фейкам та згуртування нації.

У рамках дослідження запропоновано стратегічні напрями запобігання поширенню фейків та трансформаційно-руйнівним процесам.

До таких належать: посилення виховної функції держави, формування національної ідеї, вдосконалення інформаційної політики, впровадження інструментів моніторингу інформаційного простору та використання сучасних інформаційних технологій для аналізу й протидії дезінформації. Окрема увага приділена необхідності створення системи оцінювання економічної безпеки стратегічних підприємств, що дасть змогу виявляти приховані ризики у фінансово-економічному просторі.

Таким чином, фейки не лише впливають на свідомість окремих громадян, а й формують довгострокові загрози для сталого функціонування держави. Протидія їм має бути комплексною, міжвідомчою, з орієнтацією як на технічні, так і гуманітарні механізми впливу. Лише за умови об'єднання зусиль держави, громадянського суспільства та експертного середовища можна ефективно протистояти інформаційній агресії та забезпечити стабільний розвиток України.

Список використаних джерел:

1. Вівчар О.І. Управління економічною безпекою підприємств: соціогуманітарні контексти: монографія. Тернопіль: ФОП Паляниця В.А. 2018.474с.

2. Вівчар О.І. Теоретичні аспекти безпекознавства в умовах підприємств (фундаментальні загрози в сучасному соціогуманітарному просторі). Соціально-економічні проблеми і держава. Тернопіль. 2017. Вип. (1) 16. С. 24–31. URL: http://sepd.tntu.edu.ua/images/stories/pdf/2017/17voissp.pdf

3. Герасименко Р. Факт-чекінг для журналіста: соцмережі та «вкиди» як перевірка на професійність. URL: http://ua.ejo-online.eu/2065/етика-та-якість/факт-чекінг-для-журналіста- соцмережі?print=print

4. Горбань Н. Інформаційна війна триває. Топ-5 фейків про Львів URL: http://tvoemisto.tv/news/informatsiyna_viyna_tryvaie_top5_feykiv_pro_lvi...

5. Рекун І.І. Інституціональна зміна парадигми економічної безпеки. Інституціональний вектор економічного розвитку. Збірник наукових праць МІДМУ «КПУ». Мелітополь: Вид-во КПУ. 2019. Вип. 8 (2). С. 57-65.
6. Франчук В.І. Корпоративна безпека: теоретичні засади: монографія. Львів. держ. ун-т внутр. справ. 2019. 176 с.

2

ШТУЧНИЙ ІНТЕЛЕКТ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВОЇ ІНФОРМАЦІЇ: МІЖНАРОДНІ ТРЕНДИ І МОЖЛИВОСТІ ІМПЛЕМЕНТАЦІЇ

Дзюбановська Наталія[1], Соя Маряна[1], Ліп'яніна-Гончаренко Христина[1], Кошицький Констянтин[1], Цюпа Іван[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

2.1 Актуальність дослідження та його значущість

Сучасний світ стикається з безпрецедентним поширенням фейкової інформації, яка становить серйозну загрозу для інформаційної безпеки та суспільної стабільності. З розвитком цифрових технологій та соціальних мереж дезінформація стає дедалі складнішою для виявлення, а її масштаби зростають в геометричній прогресії. Це зумовлює необхідність розробки нових підходів і технологій для боротьби з фейковими новинами. У цьому контексті штучний інтелект (ШІ) виступає потужним інструментом, здатним забезпечити швидке та ефективне виявлення дезінформації шляхом автоматизованого аналізу текстових і мультимедійних даних.

Метою даного дослідження є аналіз сучасних міжнародних тенденцій використання ШІ для виявлення фейкової інформації, вивчення основних методів і технологій, що застосовуються для вирішення цієї задачі, а також оцінка можливостей впровадження таких рішень у системах інформаційної безпеки. Важливим аспектом роботи є вивчення методів машинного навчання, відстеження аномалій, аналізу текстів і мультимедійного

контенту, які дозволяють підвищити точність і швидкість ідентифікації фейкових повідомлень. Окрім цього, дослідження спрямоване на виявлення викликів, пов'язаних з поширенням дезінформації, та обґрунтування напрямків подальшого використання ШІ для забезпечення інформаційної безпеки.

Результати дослідження сприятимуть розробці інноваційних рішень у сфері виявлення фейкової інформації, що дозволить медіаплатформам та організаціям зменшити поширення дезінформації та забезпечити захист суспільства від негативних інформаційних впливів.

2.2 Постановка проблеми та аналіз сучасного стану досліджень

У сучасному інформаційному середовищі, де кількість доступної інформації невпинно зростає, проблема поширення фейкових новин набуває особливої актуальності. Фейкова інформація, яка може включати дезінформацію, маніпулятивні повідомлення або неправдиві новини, має потенціал впливати на громадську думку, формувати політичні настрої та навіть спричиняти соціальні конфлікти. Ця проблема стає ще гострішою в умовах глобальних викликів таких, як вибори, пандемії або військові конфлікти, коли інформація стає потужним засобом впливу.

Традиційні підходи до виявлення фейкових новин, зокрема ручна перевірка фактів, виявилися недостатньо ефективними в умовах сучасних викликів через великий обсяг даних, який потребує обробки. Це зумовило необхідність впровадження новітніх технологій таких, як штучний інтелект (ШІ), для автоматизації процесу ідентифікації цього виду інформації.

Попри наявність різноманітних алгоритмів і систем на основі штучного інтелекту, проблема виявлення фейкової інформації залишається актуальною через низку викликів: недостатню точність алгоритмів, необхідність постійного навчання моделей для адаптації до нових форматів

дезінформації, ризики упередженості алгоритмів, що можуть призводити до помилкового маркування контенту, а також низький рівень медіаграмотності населення, який ускладнює критичне сприйняття інформації та підвищує вразливість до фейків.

Враховуючи ці аспекти, не менш важливим є аналіз міжнародного досвіду застосування ШІ для виявлення фейкової інформації, а також оцінка можливостей адаптації та впровадження цих технологій у практичну діяльність.

Останні кілька років характеризуються зростанням інтересу до проблеми поширення фейкових новин та можливостей застосування штучного інтелекту для їх виявлення. Швидкий розвиток цифрових технологій значно полегшив процес поширення інформації, але водночас ускладнив перевірку фактів. Для подолання цих викликів потрібні нові підходи, що поєднують технологічні рішення з соціальними ініціативами.

Зокрема, у своїй науковій статті [1] Тищенко В. аналізує різні методи навчання нейронних мереж для виявлення фейкових новин, охоплюючи текстові, візуальні та комбіновані дані. Він розглядає використання рекурентних, згорткових і глибоких нейронних мереж, а також генеративних змагальних мереж, зосереджуючись на методах навчання з вчителем та без нього. Дослідження показує, що якість і обсяг даних, а також складність фейкових матеріалів суттєво впливають на ефективність цих інструментів. Незважаючи на значні досягнення в застосуванні методів машинного та глибокого навчання, залишається потреба в подальших дослідженнях для підвищення точності технологій у розпізнаванні складних видів фейків.

Празднiков В. О. та Сугоняк І. І. у своїй роботі [2] досліджують проблему фейкового контенту та пропонують методи його виявлення за допомогою технологій машинного навчання та аналізу текстових і

мультимедійних даних. Вони підкреслюють важливість використання репрезентативних наборів даних для навчання моделей, здатних розпізнавати фейкові новини, а також розглядають застосування глибоких нейронних мереж, які можуть автоматично аналізувати лінгвістичні та візуальні характеристики для ідентифікації фейкового контенту. Дослідники також зазначають, що нині виникає гостра потреба в співпраці науковців, розвитку відкритих джерел даних та постійному вдосконаленні методів виявлення фейків для посилення захисту інформаційного простору.

Боровик Д. та Бармак О. у [3] аналізують основні підходи до боротьби з фейковими новинами, включаючи алгоритми машинного та глибокого навчання, а також досліджують аналіз сентиментів та емоцій користувачів соціальних мереж. У своїй статті вони пропонують вдосконалення методів виявлення фейкових новин за допомогою нейронних мереж, зокрема шляхом збільшення кількості нейронів у згортковому шарі та додавання шару випадкового відключення, що сприяє підвищенню ефективності моделі.

У своїй роботі [4] Дацик С. В. розробив інструмент на основі алгоритмів штучного інтелекту для виявлення дезінформації в новинах, що розповсюджуються через соціальну мережу Facebook. Він наголошує на важливості ролі ШІ у процесі виявлення фейкових новин, аналізуючи алгоритми для розпізнавання лінгвістичних шаблонів і підкреслюючи необхідність комбінованого підходу, який об'єднує зусилля людини та технології.

Дослідження Рабчуна Д. І., Тищенка В. С. та Голобородька С. О. [5] зосереджене на розробці ефективних методів автоматизованого виявлення дезінформації із застосуванням нейронних мереж та методів обробки природної мови. Особливу увагу в роботі приділено аналізу емоційного впливу, який супроводжує дезінформаційний контент, а також виявленню

технік маніпулювання емоціями аудиторії. У результаті дослідження була створена система розпізнавання та категоризації емоційного забарвлення контенту, що дозволяє ефективно відфільтровувати шкідливі матеріали та підвищувати стійкість інформаційного середовища. Дослідження також враховує етичні аспекти використання нейронних мереж та пропонує рекомендації для захисту приватності користувачів.

Одним із ключових аспектів вивчення дезінформації є її вплив на інформаційно-психологічні атаки, що можуть становити загрозу національній безпеці. У роботі [6] Дерюгіна І. К. та Батька І. І. аналізується вплив таких атак на суспільство та державну безпеку, акцентуючи увагу на здатності дезінформації маніпулювати настроями громадян та поширювати неправдиву інформацію. Дослідження підкреслює важливість адаптації законодавчих норм для забезпечення ефективного захисту від подібних загроз. Науковці рекомендують розвивати критичне мислення серед населення та стимулювати співпрацю між державними і приватними структурами для протидії інформаційно-психологічним атакам.

Незважаючи на велику кількість досліджень, присвячених виявленню фейкових новин за допомогою штучного інтелекту, подальші дослідження мають значний потенціал не лише для підвищення ефективності існуючих методів, але й для відкриття нових підходів у боротьбі з дезінформацією. Це сприятиме інтеграції сучасних технологій та методів, що дозволить більш комплексно і гнучко реагувати на виклики, які виникають в інформаційному просторі.

Метою даного дослідження є вивчення актуальних міжнародних тенденцій використання штучного інтелекту для виявлення фейкової інформації, аналіз основних методів та технологій, що застосовуються для цього, а також оцінка потенціалу впровадження таких рішень у системи інформаційної безпеки.

2.3 Виклад основного матеріалу дослідження

Фейкова інформація, що може виникати внаслідок неперевірених фактів або помилок, часто поширюється ненавмисно. Проте, коли неправдива інформація розповсюджується свідомо з метою маніпуляції думками або введення людей в оману, вона стає дезінформацією.

Дезінформація поширюється тими ж каналами, що й звичайна інформація: через телебачення, радіо, інтернет та друковані видання. Під час виборчих кампаній активно використовуються друковані матеріали, такі як політичні буклети та рекламні листівки, які, завдяки безкоштовному розповсюдженню, можуть впливати на певні соціальні групи. Радіо залишається ефективним інструментом для розповсюдження дезінформації, особливо у регіонах з обмеженим доступом до телебачення або інтернету. Телебачення також є потужним засобом поширення неправдивих новин завдяки широкому охопленню та високому рівню довіри з боку глядачів. Проте, інтернет та соціальні мережі стали головними платформами для швидкого та масового розповсюдження дезінформації, оскільки на цих платформах відсутній редакційний контроль і будь-який користувач може публікувати контент.

16 червня 2022 року Європейський Союз оновив Кодекс практики щодо боротьби з дезінформацією [7]. Учасники нового Кодексу взяли на себе зобов'язання діяти в таких напрямках: зменшення прибутків від дезінформації, підвищення прозорості політичної реклами, забезпечення цілісності наданих послуг, розширення прав користувачів, посилення ролі фактчекерів та підтримка прав дослідників. Крім того, було створено Центр прозорості та Цільову групу для моніторингу виконання положень Кодексу. Особливу увагу серед нововведень приділено активній участі приватних компаній у дотриманні вимог Кодексу. Станом на червень 2022 року до списку компаній, які повністю або частково прийняли умови

Посиленого кодексу та зобов'язалися вжити конкретних заходів для боротьби з дезінформацією на своїх платформах, увійшли: Adobe, Avaaz, Clubhouse, Crisp, Demagog, DOT Europe, European Association of Communication Agencies (EACA), Faktograf, Globsec, Google, Interactive Advertising Bureau (IAB) Europe, Kinzen, Kreativitet & Kommunikation, Logically, Maldita.es, MediaMath, Meta, Microsoft, Neeva, Newsback, NewsGuard, PagellaPoltica, Reporters without Borders (RSF), Seznam, The Bright App, The GARM Initiative, TikTok, Twitch, Twitter, Vimeo, VOST Europe, WhoTargetsMe, World Federation of Advertisers (WFA).

Розробники постійно вдосконалюють методи виявлення фейкової інформації та боротьби з дезінформацією в соціальних мережах. Серед найефективніших підходів виділяють [8] (Рисунок 2.1):

- Аналіз текстового контенту, який включає використання алгоритмів для перевірки дописів. Ці алгоритми можуть аналізувати стиль письма та лексичну схожість із відомими джерелами, що дозволяє виявляти фальсифікації. Крім того, перевіряється контекст і якщо фото або повідомлення відрізняються від типової поведінки користувача, вони позначаються як підозрілі.
- Машинне навчання, яке дає змогу створювати моделі, що навчаються на великих масивах даних. Це допомагає визначати патерни, характерні для дезінформації, причому офіційно визнані фейки використовуються як навчальні приклади для навчання штучного інтелекту.
- Відстеження аномалій, спрямоване на виявлення нетипових подій і різких змін у активності користувачів. Наприклад, якщо певний допис або зображення швидко привертає значну увагу або отримує надмірно велику кількість реакцій за короткий час,

це може свідчити про те, що інформація є фейковою або маніпулятивною.

- Аналіз мультимедійного контенту, який використовує алгоритми для виявлення ознак редагування фото та відео. Це включає оцінку технічних характеристик таких, як показники акселерометра смартфона, що дозволяють визначити, чи було зображення створене за допомогою програми. Також застосовуються аналізатори голосу та тексту, які можуть сигналізувати про підозрілу інформацію в аудіо- та відеоматеріалах.
- Виявлення ботів, яке базується на аналізі різноманітних критеріїв і параметрів таких, як частота публікацій на день, співвідношення активності акаунтів у будні та вихідні дні, а також детальний аналіз тем, слів та інших характеристик поведінки в мережі. Це допомагає визначити, чи за створенням і розповсюдженням контенту стоять спеціально розроблені програми.

Розробка та вдосконалення цих методів спрямовані на підвищення ефективності протидії фейковим новинам та дезінформації в соціальних мережах. Однак, їх об'єднання в єдину систему створює комплексний підхід для боротьби з дезінформацією на різних рівнях та у різних форматах.

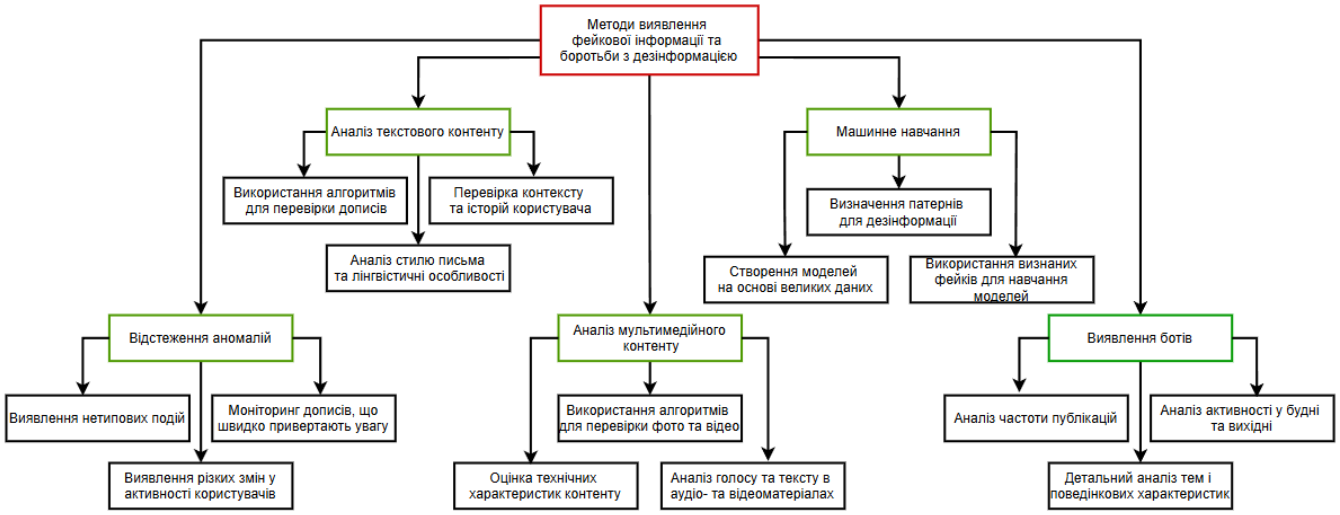


Рисунок 2.1. Методи виявлення фейкової інформації та боротьби з дезінформацією

Таким чином, виникає потреба в чіткому розумінні технологічної сутності поняття «система виявлення фейкової інформації на основі ШІ». Така система об'єднує кілька рівнів аналізу, що дозволяють автоматично ідентифікувати потенційні загрози, здійснювати аналіз контенту та поведінки користувачів, забезпечуючи комплексну оцінку даних для виявлення дезінформації.

З огляду на сучасні тенденції зростання обсягів дезінформації, значення розглянутих методів продовжує зростати, що забезпечує більш ефективний захист інформаційного простору. На глобальному рівні зростає увага до проблем, пов'язаних із дезінформацією, що стимулює міжнародну співпрацю задля розробки стандартів і практик, які сприятимуть забезпеченню прозорості та надійності інформації.

Провідні міжнародні компанії, такі як Google, Facebook (Meta), Twitter (X) та інші великі технологічні корпорації, активно використовують штучний інтелект для боротьби з фейковими новинами. Це обумовлено тим, що дезінформація стала серйозною загрозою в цифрову епоху, коли швидкість її поширення значно перевищує можливості ручної перевірки.

Facebook (Meta) застосовує подібні підходи, зокрема технології машинного навчання для автоматичного виявлення фейкових новин та модерації контенту [9]. Компанія створила комплексну систему для виявлення та позначення неправдивої інформації, а також зменшення її видимості у стрічках новин користувачів. Для цього платформа співпрацює з перевіреними фактчекерами та організаціями, що спеціалізуються на аналізі новин.

У звіті за 2021 рік Гай Розен (Guy Rosen), керівник відділу безпеки інформації (Chief Information Security Officer, CISO) компанії Meta, наголошує на рішучій позиції компанії у боротьбі з фейковими акаунтами та дезінформацією в соціальних мережах [9]. В рамках ініціативи з очищення платформи було заблоковано понад 1,3 мільярда фальшивих облікових записів та ліквідовано більше 100 мереж, що займалися скоординованою неавтентичною поведінкою. Щоб обмежити поширення неправдивої інформації, Meta зменшує видимість контенту, який позначено як неправдивий, та використовує попереджувальні позначки. Під час пандемії COVID-19 компанія видалила понад 12 мільйонів матеріалів, пов'язаних із дезінформацією про вірус та вакцини. Водночас, попри всі зусилля, Гай Розен визнає, що повністю усунути дезінформацію неможливо. Однак керівництво Meta продовжує інвестувати у нові технології та команди для вдосконалення своєї стратегії боротьби з цією проблемою. Зміни в алгоритмах і принципах роботи платформи демонструють прагнення компанії запобігти поширенню фейків та забезпечити надійність контенту, надаючи користувачам достовірну інформацію.

Інша соціальна мережа, Twitter (X) [10]., запровадила інструмент під назвою Birdwatch, який дозволяє користувачам додавати примітки до потенційно фейкових твітів, забезпечуючи додатковий контекст та

допомагаючи іншим перевіряти інформацію. Алгоритми штучного інтелекту також здійснюють аналіз контенту в режимі реального часу, щоб автоматично виявляти фейки та ненадійний контент. Кіт Коулман (Keith Coleman), віцепрезидент Twitter, активно працює над питанням дезінформації та методами її подолання на соціальних платформах. У своєму дописі про Birdwatch він зазначив, що цей інструмент отримав позитивний відгук користувачів, що підтверджується результатами понад 100 інтерв'ю з людьми різних політичних поглядів.

З 2021 року працює Центр безпеки контенту Google (Google Safety Engineering Center), який зосереджується на протидії незаконному та шкідливому контенту [11] та розробляє правила для заборони шкідливого вмісту в таких продуктах, як YouTube та Google Ads. Наприклад, протягом перших 18 місяців пандемії YouTube видалив понад мільйон відео, що містили дезінформацію про COVID-19. Співпраця з надійними партнерами, такими як Всесвітня організація охорони здоров’я (ВООЗ) та Центр з контролю та профілактики захворювань США (CDC), стала ключовою у зборі достовірної інформації.

У 2021 році Google виділив 25 мільйонів євро на створення Європейського медіа та інформаційного фонду, який підтримує медіаграмотність та фактчекінг. З березня 2020 по березень 2021 року на платформі Google було проведено понад 50 тисяч нових перевірок фактів. У листопаді 2022 року Google та YouTube надали грант у розмірі 13,2 мільйона доларів Міжнародній мережі перевірки фактів для підтримки 135 організацій (Рисунок 2.2).

Боротьба з дезінформацією залишається постійним викликом. Підрозділ Google Jigsaw вивчає нові підходи, такі як «prebunking» (попереднє розвінчування), що допомагає користувачам розпізнавати оманливі наративи. Цей підхід використовує попередження про маніпуляції

та неправдиві твердження, сприяючи підвищенню психологічної стійкості користувачів.

В Україні [12] також проводяться заходи з підвищення медіаграмотності. Так, у жовтні 2024 року Міністерство культури та інформаційної політики України запропонувало громадянам пройти тест на медіаграмотність, щоб перевірити свою здатність розрізняти фейкову інформацію. Крім того, у червні 2024 року було оприлюднено Стратегію розвитку медіаграмотності [13] до 2026 року, яка визначає медіаграмотність як компетентність, що потребує розвитку протягом усього життя.

Рисунок 2.2. Інвестиції Google у медіаграмотність та фактчекінг

Також, у жовтні 2024 року в Києві відбулася конференція [14] для локальних медіа «Фактчек у регіонах: виклики, труднощі та перспективи», організована VoxCheck за підтримки Google News Initiative, яка зібрала понад сто учасників з різних регіонів України.

Інші міжнародні технологічні компанії також розробляють подібні рішення, спрямовані на зменшення впливу дезінформації. Вони активно

використовують моделі глибокого навчання, обробку природної мови (NLP) та технології кластеризації для виявлення закономірностей у фейкових новинах та спробах маніпуляції. Наприклад, такі компанії аналізують великі масиви даних із соціальних мереж та новинних платформ, щоб ідентифікувати та зупинити поширення неправдивої інформації. У цьому контексті технологічні інновації та співпраця з партнерами стають ключовими інструментами у протидії дезінформації в сучасному світі.

2.4 Аналіз ефективності методів виявлення фейкової інформації за допомогою машинного навчання

Аналіз ефективності різних підходів (Рисунок 2.3) до виявлення фейкової інформації на основі штучного інтелекту є важливим етапом у розумінні їхніх можливостей та обмежень.

Зокрема, ключовими аспектами оцінки є точність, швидкість обробки даних, адаптивність до нових типів дезінформації та здатність до масштабування.

1. Точність моделей класифікації на основі машинного навчання

Для оцінки ефективності [15] моделей у задачах класифікації застосовуються ключові метрики, такі як точність (Precision), відновлення (Recall), F1-середнє (F1 Score), загальна точність (Accuracy) та матриця помилок (Confusion Matrix). Ці метрики дозволяють глибше аналізувати якість роботи моделі, ідентифікувати потенційні слабкі місця та оптимізувати модель для кращих результатів.

Точність (Precision) визначається як відношення кількості правильно ідентифікованих позитивних результатів до загальної кількості результатів, які модель класифікувала як позитивні. Математично вона представляється як

$$P = \frac{TP}{TP + FP},$$

де TP — true positives, а FP — false positives.

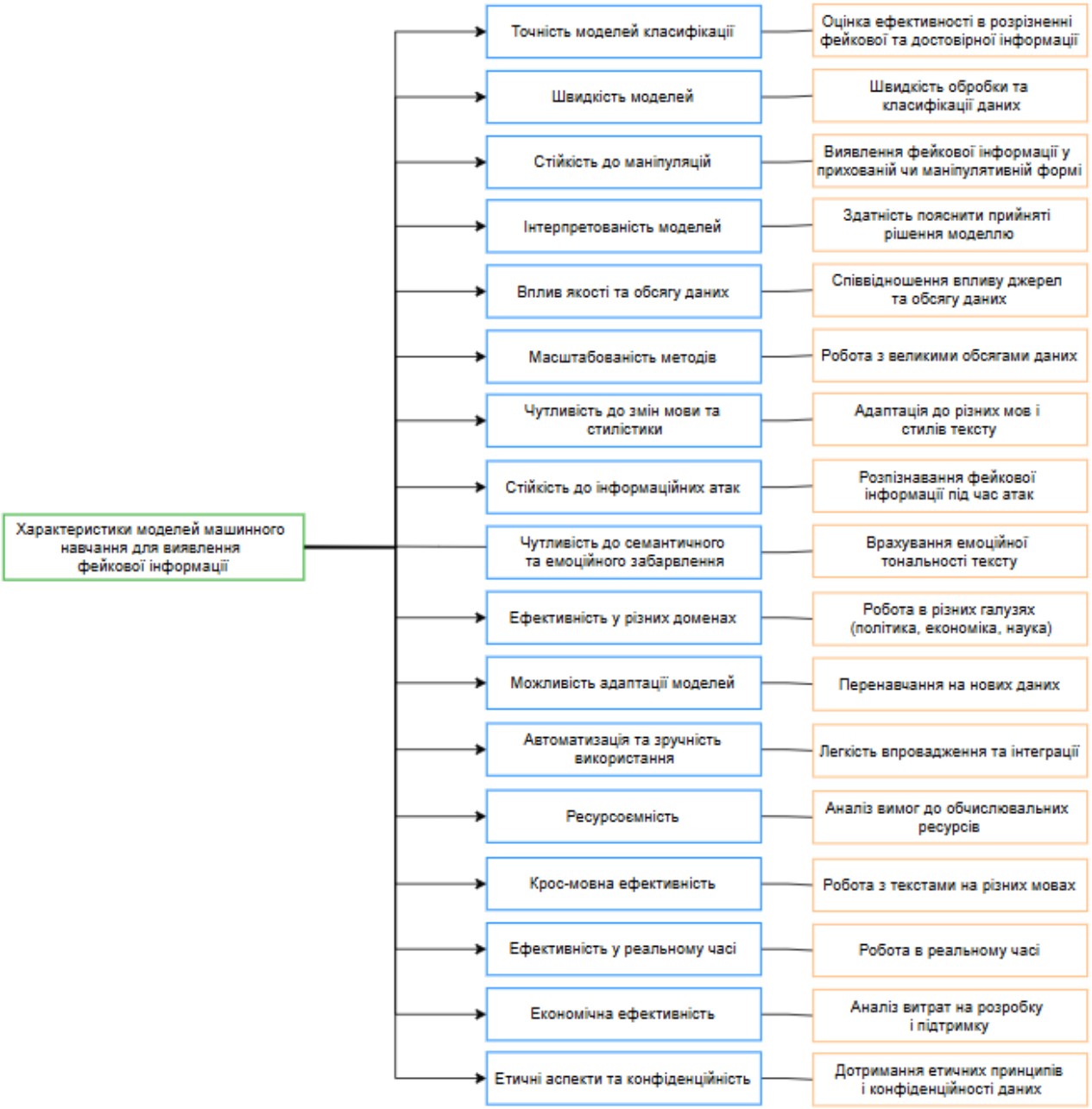


Рисунок 2.3. Ключові характеристики моделей машинного навчання для боротьби з дезінформацією

Відновлення (Recall) вимірює здатність моделі ідентифікувати всі реальні позитивні випадки в датасеті. Вона визначається як відношення кількості правильно ідентифікованих позитивних результатів до суми

правильно ідентифікованих позитивних результатів та результатів, які є позитивними, але були пропущені моделлю. Формула для розрахунку:

$$R = \frac{TP}{TP+FN},$$

де FN — false negatives.

F1-середнє (F1 Score) є гармонійним середнім між точністю та повнотою, забезпечуючи баланс між цими двома метриками. Це особливо корисно в ситуаціях, де дисбаланс між класами може спричинити перекоси в одній з метрик на користь іншої. F1 визначається як

$$F1 = 2 \cdot \frac{P \cdot R}{P+R}.$$

Загальна точність (Accuracy) вимірює відсоток випадків, правильно класифікованих моделлю, та визначається формулою

$$A = \frac{TP+TN}{TP+TN+FP+FN},$$

де TN — true negatives.

Матриця помилок (Confusion Matrix) надає візуалізацію результатів класифікації, представляючи кількість TP, TN, FP, та FN у формі матриці. Це дозволяє не лише визначити точність моделі, але й зрозуміти типи помилок, які робить модель.

2. Швидкість моделей машинного навчання

Швидкість роботи моделей машинного навчання є критичним фактором у задачах, де необхідно миттєво реагувати на вхідні дані. Це може бути важливо в системах реального часу таких, як виявлення шахрайства, обробка зображень або аналіз текстових потоків. Основні метрики, що використовуються для оцінки швидкості, включають:

Час інференції (Inference Time), визначає, скільки часу займає модель для обробки одного зразка даних та надання передбачення. Це важливо для оцінки, наскільки швидко модель може відповідати на запити в реальному часі. Математично час інференції визначається як:

$$Inference\ Time = \frac{\text{Total Processing Time}}{\text{Number of Predictions}}$$

Вища швидкість інференції є бажаною, особливо для інтерактивних систем та мобільних додатків, де затримка може значно впливати на користувацький досвід.

Пропускна здатність (Throughput) показує, скільки запитів або даних модель може обробити за одиницю часу. Вона визначається як кількість передбачень, які модель може виконати за фіксований проміжок часу. Формула для розрахунку пропускної здатності:

$$Throughput = \frac{\text{Number of Predictions}}{\text{Time Period}}$$

Висока пропускна здатність є важливою для великих систем, де обробляються великі обсяги даних, наприклад, у випадках аналізу стримінгових відео або аудіо в режимі реального часу.

Затримка (Latency) вимірює час від моменту надходження запиту до моменту отримання результату від моделі. Це включає час обробки вхідних даних, виконання передбачення та передачу результату. Формула для розрахунку затримки:

$$Latency = \text{Time to First Byte (TTFB)} + \text{Processing Time}$$

Низька затримка є важливою у критичних додатках таких, як автономне водіння, системи безпеки, де навіть невеликі затримки можуть призвести до серйозних наслідків.

3. Витрати на обчислювальні ресурси

Оцінює ефективність використання ресурсів таких, як обсяг пам'яті та використання процесорного часу. Може бути оцінено через затрати енергії або кількість використаних обчислювальних одиниць:

$$Cresources = \alpha \cdot Ttrain + \beta \cdot Tinference$$

де α та β – вагові коефіцієнти залежно від важливості ресурсів у конкретному застосуванні.

2.5 Висновки з проведеного дослідження

Дослідження дезінформації та фейкових новин підкреслює складність та багатогранність цієї проблеми, що вимагає комплексного підходу для її вирішення. Дезінформація становить серйозний виклик для сучасного суспільства, оскільки здатна маніпулювати громадською думкою та впливати на соціально-політичні процеси. Застосування новітніх технологій, зокрема штучного інтелекту, відіграє ключову роль у протидії дезінформації, забезпечуючи швидке та ефективне виявлення неправдивої інформації, аналіз контенту та поведінки користувачів, а також автоматизацію процесів модерації.

Попри значні зусилля технологічних компаній у сфері виявлення фейкових новин, ця проблема залишається актуальною, що вимагає подальшого вдосконалення методів моніторингу та виявлення дезінформації. Це включає розвиток алгоритмів машинного навчання, здатних розпізнавати патерни фальсифікацій та аномальну активність користувачів, а також аналіз мультимедійного контенту для відстеження змін у формі та якості інформації.

Ключовим аспектом є співпраця між технологічними компаніями, фактчекерами, міжнародними організаціями та державними урядами. Така взаємодія сприяє розробці стандартів і практик, які підвищують прозорість та надійність інформації в цифрову епоху. Водночас важливо приділяти увагу розвитку критичного мислення серед громадян, що дозволить їм самостійно розрізняти надійні та недостовірні джерела інформації.

Загалом, успішна боротьба з дезінформацією потребує поєднання технологічних інновацій, дотримання етичних норм та соціальної відповідальності. Інтеграція комплексних підходів до виявлення та моніторингу фейкових новин може значно зменшити їхній вплив на суспільство, забезпечуючи якісніше інформаційне середовище.

Запровадження заходів для боротьби з дезінформацією та фейковими новинами значно розширить можливості у сфері інформаційної безпеки, громадської освіти та технологічного розвитку, суттєво підвищуючи ефективність цих зусиль. Необхідно продовжувати впровадження технологій штучного інтелекту та машинного навчання, які здатні автоматизувати процеси виявлення та оцінки інформації. Це дозволить значно скоротити час реагування на фейкові новини та підвищити точність їх виявлення. Компанії можуть інтегрувати алгоритми для аналізу тексту, зображень та відео, які виявляють аномалії або патерни, що вказують на маніпуляцію інформацією. Крім того, важливо налагодити тісну співпрацю між технологічними компаніями, державними органами та громадськими організаціями. Створення міжвідомчих платформ для обміну даними та кращими практиками може сприяти формуванню єдиного фронту в боротьбі з дезінформацією. Такі платформи можуть включати механізми для спільної перевірки фактів, розвитку відкритих даних та ресурсів для навчання користувачів.

Важливою ініціативою є впровадження освітніх програм на всіх рівнях, від шкіл до університетів, які сприятимуть розвитку критичного мислення. Навчання громадян розпізнавати фейкові новини, розуміти механізми дезінформації та користуватися надійними джерелами є ключовим елементом для формування стійкого суспільства.

Крім того, необхідне ефективне правове регулювання для роботи платформ обміну інформацією, зокрема встановлення чіткої відповідальності за розповсюдження фейків. Важливим аспектом цього процесу є забезпечення прозорості алгоритмів, які використовуються для ранжування контенту, адже це підвищить довіру користувачів до інформаційних платформ та знизить ризик маніпуляцій інформацією.

Загалом впровадження заходів для протидії фейковій інформації має значний потенціал. Комбінація технологічних рішень, освітніх програм та правових ініціатив сприятиме суттєвим позитивним змінам у створенні якісного інформаційного середовища та підвищенню рівня обізнаності громадян.

Список використаних джерел:

1. Тищенко В. Аналіз методів навчання та інструментів нейромереж для виявлення фейків. Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка». (2023). 4(20), 20–34. https://doi.org/10.28925/2663-4023.2023.20.2034.
2. Праздніков, В. О., Сугоняк, І. І. Моделі та методи машинного навчання для розпізнавання фейкового контенту. Технічна інженерія. 2023. № 2 (92). С. 131–136. URL: http://eztuir.ztu.edu.ua/123456789/8391. DOI: https://doi.org/10.26642/ten-2023-2(92)-131-136.
3. Боровик Д., Бармак О. Удосконалений метод виявлення фейкових новин на основі використання CNN нейромережі. Вісник Хмельницького національного університету. 2023. Том 2, №5. С. 19–24. DOI: 10.31891/2307-5732-2023-327-5-19-24.
4. Дацик С. В. Використання штучного інтелекту для виявлення дезінформації в новинах соціальної мережі Facebook : кваліфікаційна робота на здобуття освітнього ступеня магістр за спеціальністю “122 – комп’ютерні науки” / С. В. Дацик. – Тернопіль: ТНТУ, 2024. – 90 с.
5. Рабчун Д. І., Тищенко В. С., Голобородько С. О. Ефективне розпізнавання дезінформації за допомогою нейронних мереж: фокус на виявлення емоційного впливу. Елекомунікаційні та інформаційні технології. 2024. № 2 (83). С. 37–48. DOI: 10.31673/2412-4338.2024.024860.

6. Дерюгін І.К., Батько І.І Інформаційно-психологічні атаки як загроза державній безпеці. Електронне наукове видання «Аналітично-порівняльне правознавство». 2023. № 05. С. 315–319. DOI https://doi.org/10.24144/2788-6018.2023.05.57.
7. Оновлення кодексу ЄС щодо боротьби з дезінформацією: основні положення. URI: https://jurliga.ligazakon.net/news/212519_onovlennya-kodeksu-s-shchodo-borotbi-z-deznformatsyu-osnovn-polozhennya.
8. Сучасні методи виявлення фейків у соціальних мережах. URI: https://optima.study/blog/suchasni-metodi-viyavlennya-feykiv-u-socialnih-merezhah.
9. Rosen Guy. How We're Tackling Misinformation Across Our Apps. URI: https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/.
10. Coleman Keith. Introducing Birdwatch, a community-based approach to misinformation. URL: https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.
11. Google's approach to fighting misinformation online. URL: https://safety.google/intl/en_uk/stories/fighting-misinformation-online/.
12. МОН України https://mon.gov.ua/news/testuimediahramotnist-ukraintsiv-zaproshuiut-sklasty-test-na-vminnia-rozrizniaty-feiky.
13. https://ms.detector.media/internet/post/35133/2024-06-05-strategiyu-z-rozvytku-mediagramotnosti-do-2026-roku-oprylyudnyly-dlya-gromadskosti/.
14. https://ms.detector.media/onlain-media/post/36315/2024-10-02-faktchek-u-regionakh-dosvid-lokalnykh-media/.

15. Xing, H. J., Liu, W. T., & Wang, X. Z. (2024). Bounded exponential loss function based AdaBoost ensemble of OCSVMs. Pattern Recognition, 148, 110191.

3

ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ В УКРАЇНСЬКИХ ТЕКСТОВИХ ДАНИХ

Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Юрків Христина[1], Івасечко Андрій[1], Галяс Юрій[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

3.1 Огляд існуючих рішень

У сучасному інформаційному просторі величезна кількість даних генерується щоденно, серед яких значну частку становить новинний контент. У контексті зростаючої глобалізації та доступності інформаційних технологій, інформація швидко поширюється через мережу, що робить її потужним інструментом впливу на громадську думку. Однак, це також відкриває шлях для масового розповсюдження дезінформації, яка може мати значні наслідки для суспільства, політики та міжнародних відносин. Розібратися в цьому потоці інформації та відрізнити достовірні дані від фальшивих стає все більш важливим завданням.

Повномасштабне вторгнення росії в Україну в лютому 2022 року стало яскравим прикладом використання дезінформації, як зброї в умовах гібридної війни. Це викликало необхідність у розробці ефективних інструментів для аналізу та класифікації інформаційного контенту, з метою ідентифікації та протидії дезінформації.

У цьому контексті, машинне навчання та методи обробки природної мови відіграють ключову роль у виявленні та аналізі фальшивих новин. Застосування цих технологій дозволяє автоматизувати процес виявлення дезінформації, забезпечуючи швидку та ефективну обробку великих обсягів даних. Водночас, розробка ефективних моделей машинного навчання для класифікації інформації вимагає глибокого розуміння специфіки даних, методів попередньої обробки та оптимізації моделей.

Ця стаття присвячена аналізу датасету, що містить заголовки новин, зібрані протягом російсько-українського конфлікту, з метою ідентифікації дезінформації. Розгляд застосування різноманітних методів машинного навчання, включаючи логістичну регресію, опорні вектори (SVM), випадковий ліс, градієнтний бустінг, KNN, дерева рішень, XGBoost та AdaBoost для класифікації текстових даних. Заключний розділ присвячений оцінці ефективності цих моделей з використанням стандартних метрик оцінки, включаючи точність, повноту, F1-середнє, загальну точність та матрицю помилок, що дозволяє визначити найбільш ефективні методи для боротьби з дезінформацією в умовах інформаційної війни.

В області інформаційної безпеки та аналізу медіаконтенту, важливість виявлення та аналізу дезінформації, особливо антивакцинного контенту на платформах соціальних мереж, як Twitter, набуває особливої актуальності. У [1] розробляє мовно-нейтральні моделі для виявлення такого контенту на велику шкалу, використовуючи багатогранні представлення повідомлень у мережах. Водночас, [2] акцентують на викликах, пов'язаних із попереднім навчанням графових нейронних мереж для контекстно-орієнтованого виявлення фейкових новин, вказуючи на стратегічні та ресурсні обмеження. У [3] проводять порівняльний аналіз навчених під наглядом та без нагляду машинних алгоритмів навчання для

виявлення фейкових новин, демонструючи їхню продуктивність, ефективність та міцність.

Дослідження, що охоплюють аналіз дезінформації та громадської думки щодо російсько-української війни, використовують різноманітні методології, включаючи аналіз настроїв, створення та аналіз датасетів, а також дослідження впливу війни на вибір мови. Одне дослідження моделює та кластеризує тенденції настроїв різних країн щодо війни [4], тоді як інше пропонує детальний датасет твітів, пов'язаних з кризою [5]. Також аналізується дискурс у Twitter щодо війни, з особливим фокусом на мову та географічне походження твітів [6]. Інше дослідження зосереджено на викликах маркування чутливого контенту та його психологічному впливі на анотаторів [7]. Вивчається також, як війна впливає на вибір мови українців у Twitter, аналізуючи зміни в мовних перевагах з часом [8]. Окреме дослідження надає уявлення про активність на subreddit'ax, пов'язаних з конфліктом, аналізуючи обсяги постів та коментарів та рівень залучення [9]. Дослідження [10] демонструє, що застосування моделі BERT для виявлення фейкових новин досягає точності 79.88% та площі під кривою ROC 0.87, підкреслюючи її потенціал у боротьбі з дезінформацією у соціальних мережах.

Як підтверджено вищезазначеним аналізом, дослідження в області виявлення дезінформації, зокрема фейкових новин, є значущим напрямком у контексті інформаційної безпеки та аналізу медіаконтенту. Особливу увагу привертає потреба в розробці та застосуванні передових технологічних рішень, для ефективного виявлення фейкових новин. Проте, існує помітна відсутність досліджень, спрямованих на аналіз дезінформації українською мовою, що ставить перед науковою спільнотою завдання розширення мовного спектру досліджень у цій галузі. Враховуючи це, метою даного дослідження є розробка інтелектуальних методів виявлення

дезінформації, зокрема фейкових новин, з акцентом на україномовний контент, що дозволить підвищити рівень інформаційної безпеки та гарантувати інтегрітет новинного простору в умовах сучасного інформаційного суспільства.

3.2 Методологія дослідження

3.2.1 Опис датасету

Для збору та обробки даних використовувався датасет [11], що містить приблизно 10700 заголовків новин про російсько-українську війну, зібраних з 24 лютого по 11 грудня 2022 року, що охоплює період з початку повномасштабного вторгнення.

Набір даних для аналізу дезінформації в українському медіапросторі складається з двох основних файлів: "data_set_4.csv" (Таблиця 3.1) та "news_data.csv" (Таблиця 3.2), кожен з яких містить новинні заголовки, класифіковані як правдиві ("True") або фальшиві ("False"). У файлі "data_set_4.csv" зареєстровано 8,237 правдивих та 2,498 фальшивих новин, тоді як "news_data.csv" містить значно більшу кількість правдивих новин — 48,006, проти 2,024 фальшивих, відображаючи широкий спектр інформаційного контенту, зібраного для дослідження розповсюдження дезінформації протягом російсько-української війни.

Таблиця 3.1.

Структура data_set_4.csv

Поле	Опис	Тип Даних
Text	Заголовок новини	Текст
Label	Мітка, що позначає новину як правдиву ("True") чи фальшиву ("False")	Булеве
Link	Посилання на джерело новини (доступно лише в цьому файлі)	Текст

Таблиця 3.2.

Структура news_data.csv

Поле	Опис	Тип Даних
Text	Заголовок новини	Текст
Label	Класифікація новини як правдива ("True") чи фальшива ("False")	Булеве

Обидва датасети (табл.3.1, 3.2) використовуються для аналізу дезінформації в українському медіапросторі та включають дані, які були зібрані з офіційних та неофіційних джерел з метою дослідження розповсюдження фейкових новин в контексті російсько-української війни.

На графіках розподілу міток видно, що в обох наборах даних переважає кількість істинних повідомлень над фальшивими. Наприклад, у першому наборі даних співвідношення істинних повідомлень до фальшивих є високим, що свідчить про зосередження на достовірній інформації. Рисунок 3.1.

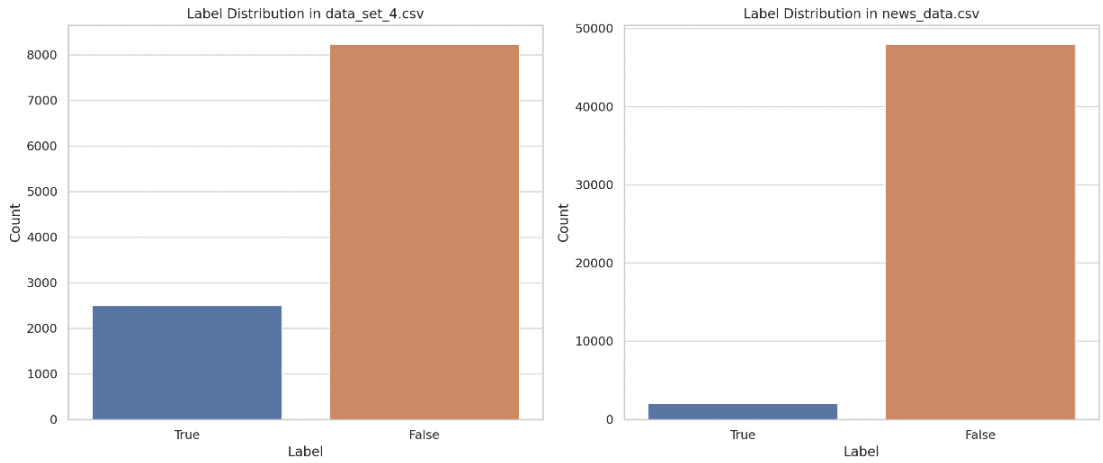


Рисунок 3.1. Розподіл міток

Аналіз найчастіше вживаних слів у першому наборі даних виявив слова, які з'являються сотні разів. Це дозволяє ідентифікувати ключові теми дискусій або новин, наприклад, слово "Україна" може зустрічатися найчастіше, підкреслюючи географічний або політичний фокус зібраних даних (Рисунок 3.2).

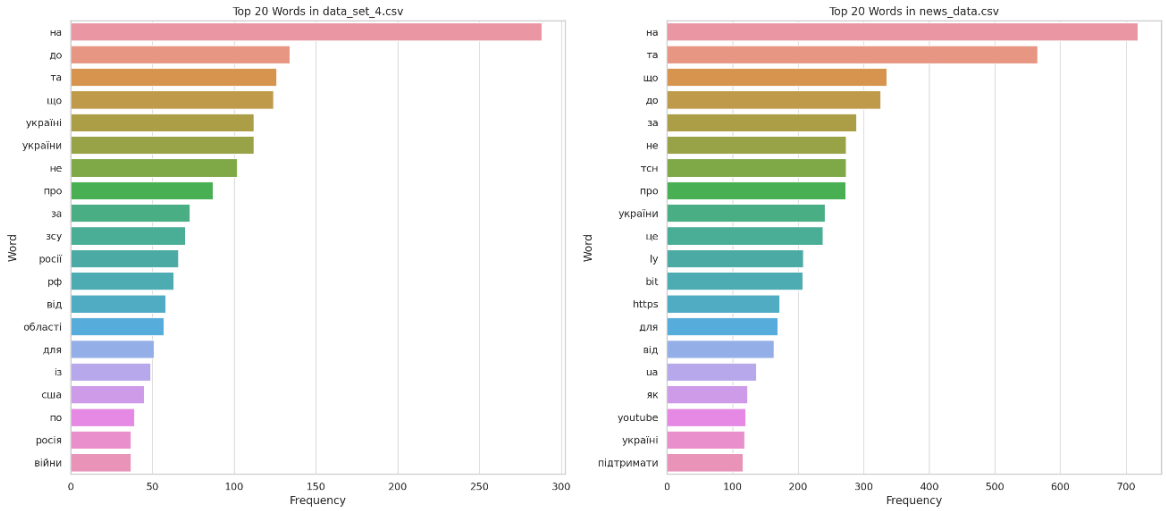


Рисунок 3.2. Топ 20 Слів

Більшість текстів у першому наборі даних мають довжину від 100 до 500 символів. Така інформація допомагає зрозуміти середній обсяг повідомлень, який може вказувати на переважання коротких новин або оглядів, Рисунок 3.3.

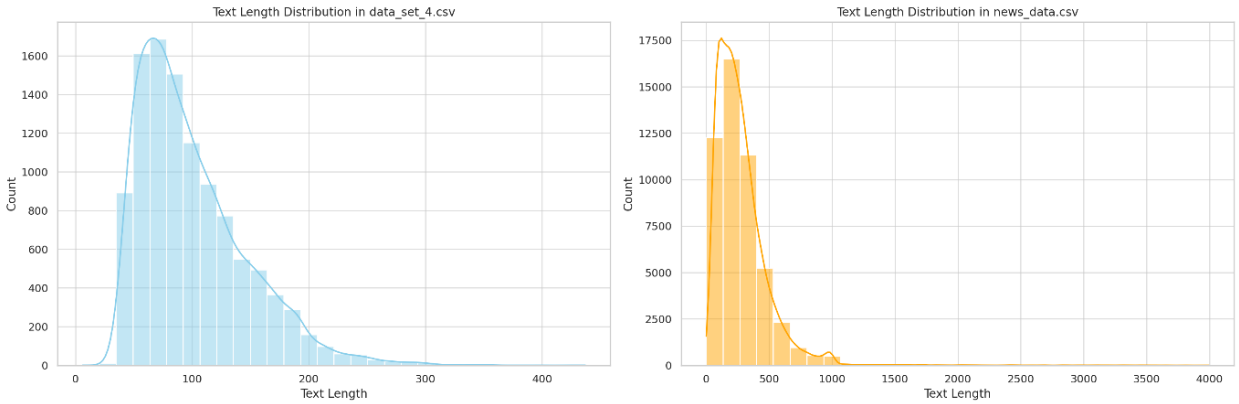


Рисунок 3.3. Розподіл довжини текстів

Аналіз представлених датасетів, що охоплюють заголовки новин про російсько-українську війну, демонструє значну перевагу достовірної інформації у порівнянні з фальшивою, що підкреслює акцент на якісному контенті. Враховуючи, що набір "news_data.csv" містить значно більше записів порівняно з "data_set_4.csv", логічно використовувати перший для тренування моделей машинного навчання, оскільки він забезпечує ширший спектр інформації та більшу кількість прикладів для навчання. Водночас,

менший набір "data_set_4.csv" може слугувати відмінним набором для тестування та перевірки ефективності моделей на меншій, але конкретній вибірці даних. Це дозволить оцінити здатність моделі генералізувати навчання на нових, раніше невідомих даних, підкреслюючи її практичну цінність у реальних умовах аналізу дезінформації.

3.2.2 Опис використаних класифікаторів

У сучасному аналізі даних, особливо в контексті виявлення дезінформації, використання алгоритмів машинного навчання для класифікації текстових даних стає ключовим інструментом для розробників та аналітиків. Різноманітність методів класифікації [12] таких, як логістична регресія, опорні вектори (SVM), випадковий ліс, градієнтний бустінг, k-найближчих сусідів (KNN), дерево рішень, XGBoost, та AdaBoost, забезпечує широкий спектр підходів для аналізу та класифікації даних. Кожен з цих алгоритмів має свої унікальні переваги та обмеження, що робить їх більш або менш придатними для конкретних типів даних та аналітичних завдань. Основною метою є вибір оптимального класифікатора, який найкраще відповідає специфіці задачі та даних, з якими працюємо.

Логістична регресія [13] є методом класифікації, що використовується для передбачення ймовірності наявності двох можливих результатів залежно від однієї або кількох незалежних змінних. Вона перетворює лінійну комбінацію вхідних даних на ймовірність за допомогою логістичної функції. Переваги логістичної регресії включають простоту інтерпретації результатів, але вона може бути обмежена при аналізі складних взаємозв'язків між змінними та вимагає припущення про лінійність зв'язків. Математично, логістична регресія моделює ймовірність

P(Y=1) як функцію від X, де Y — залежна змінна, а X — набір незалежних змінних. Ймовірність описується рівнянням:

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

де e — основа натурального логарифму, β_0 — константа (відсічення), а β_1,..., β_k — коефіцієнти незалежних змінних. Це рівняння дозволяє оцінити ймовірність того, що спостереження належить до класу 1, залежно від значень незалежних змінних.

Алгоритм опорних векторів (SVM) [14] знаходить гіперплощину в багатовимірному просторі, що найкраще відокремлює різні класи даних, максимізуючи відстань між найближчими точками даних (опорними векторами) різних класів. До переваг SVM належать висока точність у класифікаційних задачах, особливо на відносно невеликих датасетах, а також гнучкість завдяки можливості використання різних ядерних функцій. Однак, недоліками є висока вимогливість до обчислювальних ресурсів при великому обсязі даних та складність у трактуванні моделі. Математично, SVM шукає розв'язок оптимізаційної задачі: мінімізувати $\frac{1}{2} || w ||^2$ за умови, що

$$y_i \ (w \cdot x_i + b) \geq 1 \text{ для всіх i,}$$

де w — ваговий вектор гіперплощини, b — зсув, x_i— вектори ознак, а y_i — мітки класів.

Алгоритм випадкового лісу (Random Forest) [15] створює ансамбль дерев рішень, тренуючи кожне дерево на випадкових підмножинах тренувального набору даних і ознак, що забезпечує високу точність і універсальність моделі. Перевагами випадкового лісу є здатність ефективно обробляти великі датасети з високою розмірністю ознак та менша схильність до перенавчання порівняно з окремими деревами рішень. Проте серед недоліків варто відзначити високі вимоги до обчислювальних

ресурсів та складність інтерпретації моделі через велику кількість дерев. Математична інтерпретація випадкового лісу заснована на принципі "мудрості натовпу", де кінцеве рішення моделі визначається голосуванням між деревами для задач класифікації або усередненням виходів для задач регресії. Точніше, для класифікації:

$$Y = mode\{y_1, y_2, \dots, y_n\},$$

а для регресії:

$$Y = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

де y_i — прогноз кожного дерева, а Y — кінцевий прогноз ансамблю.

Градієнтний бустінг [16] — це метод ансамблевого машинного навчання, що покращує прогнози шляхом послідовного навчання слабких моделей, зазвичай дерев рішень, з метою мінімізації функції втрат. Він відрізняється високою точністю прогнозування та гнучкістю в налаштуванні параметрів, але може бути схильним до перенавчання при неправильній конфігурації та вимагає більше часу та обчислювальних ресурсів для тренування порівняно з іншими алгоритмами. Математично, градієнтний бустінг виконує оптимізацію, адаптивно зменшуючи різницю між фактичними та прогнозованими значеннями за допомогою градієнтного спуску, де оновлення моделі в m-тій ітерації визначається як:

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

де $F_{m-1}(x)$ — прогноз на попередньому кроці, $h_m(x)$— слабкий класифікатор, а α_m— швидкість навчання.

Алгоритм k-найближчих сусідів (KNN) [17] класифікує об'єкти на основі найближчих тренувальних прикладів у просторі ознак, де "k" вказує на кількість сусідів, що враховуються для визначення класу нового об'єкта. Перевагами KNN є простота реалізації та здатність ефективно працювати з мультикласовими датасетами, проте він вимагає значних обчислювальних ресурсів для зберігання тренувальних даних і визначення сусідів у великих

датасетах, а також є чутливим до несуттєвих ознак та масштабування даних. Математично, класифікація об'єкта x в алгоритмі KNN визначається більшістю голосів його сусідів, де кожен сусід вагоміше відповідно до зворотньої відстані до x, зазвичай використовуючи Евклідову відстань:

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

де x — точка для класифікації, x_i — точка з тренувального набору, j варіюється від 1 до m, кількості ознак. Клас x визначається на основі найчастішого класу серед k найближчих сусідів.

Алгоритм дерева рішень [18] будує модель прогнозування у формі деревоподібної структури, розбиваючи датасет на все дрібніші підмножини, одночасно розвиваючи асоційоване дерево рішень. Перевагами дерева рішень є легкість інтерпретації, здатність обробляти як числові, так і категоріальні дані, та непотрібність нормалізації даних. Однак, дерева рішень схильні до перенавчання, особливо при великій глибині дерева, та можуть бути нестабільними, тобто малі зміни в даних можуть призвести до значно інших дерев рішень. Математично, дерево рішень використовує поняття інформаційного прибутку або зменшення невизначеності (ентропії) для вибору атрибута, що найкраще розділяє датасет на підмножини, відповідно до цільової змінної. Інформаційний прибуток для атрибута A визначається як

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

де T — тренувальний набір, $Values(A)$ — множина всіх можливих значень атрибута A, T_v — підмножина T, для якої атрибут A має значення v, а $Entropy(T)$ — ентропія тренувального набору T.

XGBoost (eXtreme Gradient Boosting) [19] — це високоефективна реалізація алгоритму градієнтного бустінгу, який оптимізує як лінійні моделі, так і дерева рішень. Алгоритм вирізняється високою швидкістю

виконання, здатністю ефективно масштабуватися на великі датасети та вбудованою підтримкою регуляризації, що допомагає зменшити перенавчання. Однак, XGBoost може бути викликом для налаштування через велику кількість гіперпараметрів та вимагає більше часу на тренування порівняно з менш складними моделями. Математично, XGBoost мінімізує втрати використовуючи градієнтний спуск, де цільова функція включає як функцію втрат L, так і штраф за складність моделі Ω, що додає регуляризацію:

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

де y_i — істинні мітки, $\hat{y}_i$ — передбачені мітки, f_k— функції, що представляють окремі дерева, а $\Omega(f_k)$ включає терміни для регуляризації, такі як кількість листків і сума квадратів ваг вузлів, зменшуючи тим самим ризик перенавчання.

AdaBoost (Adaptive Boosting) [20] — це алгоритм машинного навчання, що поєднує кілька слабких класифікаторів для створення сильного класифікатора, використовуючи ітеративний підхід для корекції помилок попередніх класифікаторів шляхом надання більшої ваги важким для класифікації спостереженням. Перевагою AdaBoost є його здатність покращувати точність прогнозування, простота реалізації та автоматична корекція класифікаторів, що недостатньо добре працюють. Однак, він може бути схильний до перенавчання при наявності викидів або дуже шумних даних та вимагає уважного підбору кількості ітерацій. Математично, AdaBoost адаптує ваги тренувальних спостережень, w_i, підвищуючи ваги неправильно класифікованих спостережень. На кожному ітераційному кроці t, класифікатор h_t вибирається для мінімізації зваженої суми помилок. Кінцевий класифікатор визначається як вагома сума

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

де α_t — вага, надана класифікатору h_t, яка залежить від його точності.

Висновки з опису використаних класифікаторів підкреслюють важливість адаптації моделі до специфіки датасету та аналітичної задачі. Ефективність кожного методу залежить від розміру та якості даних, складності взаємозв'язків у датасеті, а також від специфічних вимог до точності та інтерпретації результатів. У контексті аналізу дезінформації, вибір між простотою інтерпретації логістичної регресії та високою точністю, але складністю налаштування XGBoost або SVM, може визначити успіх або невдачу в ідентифікації фальшивих новин. Таким чином, ретельний вибір та налаштування класифікаторів є критично важливим для розробки ефективних інструментів боротьби з дезінформацією в умовах російсько-української війни.

3.2.3 Тренування моделей

Процедура тренування моделей (Рисунок 3.4) [21] на навчальній вибірці є фундаментальним етапом у процесі розробки ефективних алгоритмів машинного навчання для класифікації текстових даних. У цьому контексті, використання датасетів "news_data.csv" та "data_set_4.csv" для тренування та тестування моделей, відповідно, забезпечує цінну основу для аналізу дезінформації. Ініціалізація процедури починається з імпортування та попередньої обробки даних, зокрема видалення записів без текстового вмісту, що забезпечує чистоту даних для наступних етапів обробки.

Векторизація тексту за допомогою TF-IDF [22] (Term Frequency-Inverse Document Frequency) є ключовим кроком у підготовці даних, оскільки вона перетворює текстові дані у числовий формат, зробивши їх придатними для обробки моделями машинного навчання. Цей метод враховує не тільки частоту слів у тексті, але й їх унікальність через інверсію

частоти документів, що дозволяє моделі краще ідентифікувати важливі ознаки в тексті.

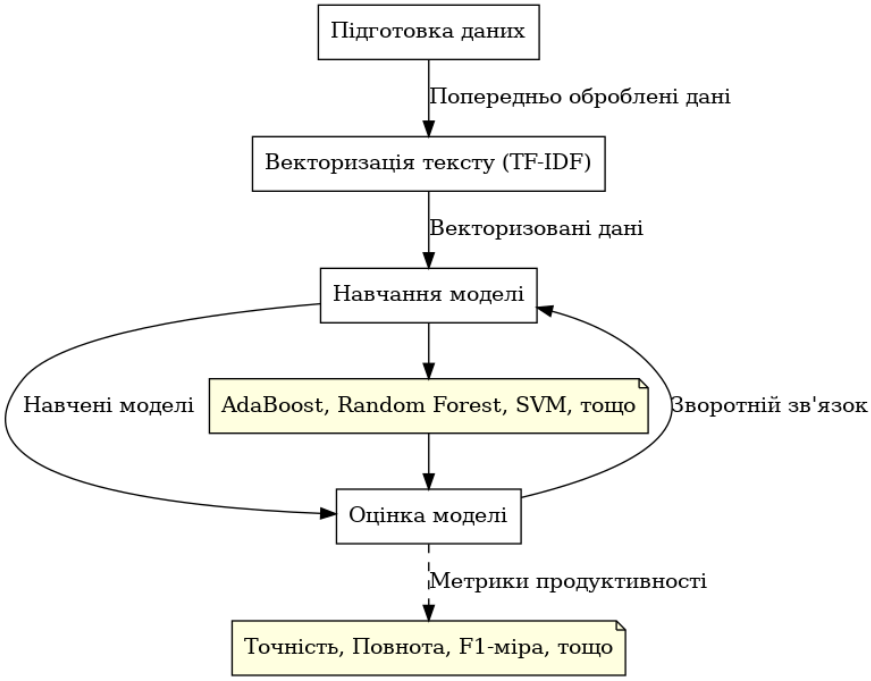


Рисунок 3.4. Процес тренування моделей

Після підготовки векторизованих навчальних і тестових даних, наступний етап включає тренування різних моделей на навчальній вибірці. Цей процес вимагає застосування алгоритмів машинного навчання, до навчального набору даних для формування моделі, здатної ефективно класифікувати текст на основі вивчених ознак. Кожна модель адаптує свої параметри таким чином, щоб максимально знизити помилки на навчальному наборі, одночасно забезпечуючи здатність до генералізації на нових даних.

Оцінювання ефективності кожної тренованої моделі на тестовому наборі даних є вирішальним для визначення її придатності та ефективності у задачах класифікації. Використання метрик оцінки таких, як точність, повнота, F1-середнє, та загальна точність, дозволяє глибоко аналізувати та порівнювати продуктивність моделей. Результати оцінювання можуть бути візуалізовані для кращого розуміння ефективності моделі, а найкраща

модель може бути вибрана для подальшого використання або збережена для майбутнього аналізу.

Таким чином, процедура тренування моделей на навчальній вибірці з використанням попередньо оброблених та векторизованих текстових даних є фундаментальною для розробки надійних інструментів машинного навчання, здатних ефективно класифікувати та аналізувати текст на предмет дезінформації.

3.3 Результати дослідження

Далі розглянемо реалізацію та оцінку ефективності різних моделей машинного навчання для класифікації даних про дезінформацію в українському медіапросторі, що є особливо актуальним в контексті повномасштабного вторгнення росії. Аналізуючи широкий спектр алгоритмів, прагнемо виявити найбільш ефективні методи для точного виявлення та розрізнення інцидентів дезінформації. Цей аналіз ґрунтується на комплексному порівнянні загальної точності класифікації, а також специфічних метрик таких, як точність, відновлення (recall), та F1-оцінка для кожного з класів.

Модель логістичної регресії показала (Рисунок 3.5) загальну точність класифікації 86.4% на тестовій вибірці з 10735 прикладів, демонструючи високу ефективність у визначенні істинних значень з F1-оцінкою 0.92 для класу "True". Однак, модель виявилася менш точною у визначенні класу "False", з точністю прогнозування 0.96 та низьким показником відновлення (recall) 0.44, що свідчить про значну кількість помилок другого роду, як це відображено в матриці помилок з 1407 неправильно класифікованими прикладами.

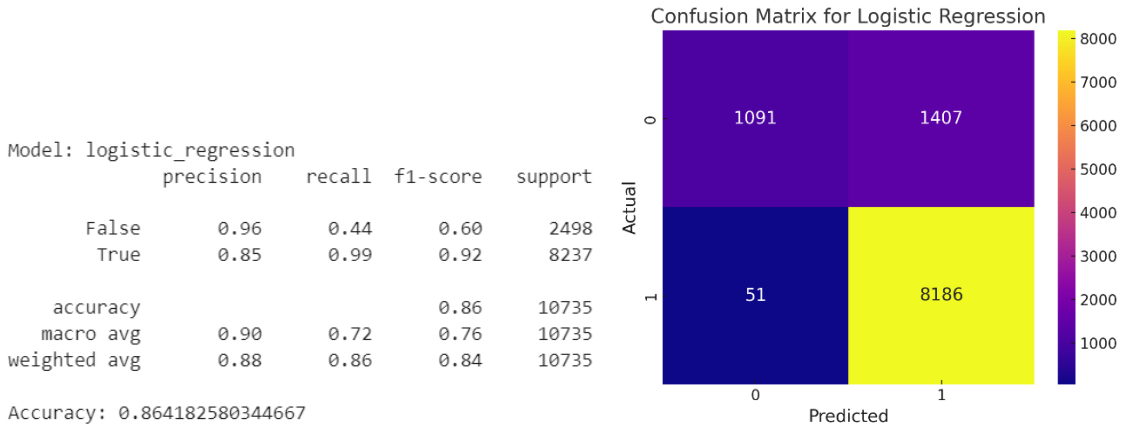


Рисунок 3.5. Результати LogisticRegression

Модель SVM показала (Рисунок 3.6) високу загальну точність класифікації на рівні 93.6% для тестової вибірки, вказуючи на її ефективність у розрізненні між класами "True" та "False". Специфічні показники точності та відновлення (recall) для класу "False" становили 0.97 та 0.75 відповідно, що демонструє здатність моделі добре ідентифікувати негативні випадки, хоча з деякою кількістю помилок. Водночас, високі показники для класу "True" з точністю 0.93 та відновленням 0.99 свідчать про мінімальну кількість помилково негативних класифікацій, підтверджені низькою кількістю помилок у матриці помилок (618 для "False" та 68 для "True").

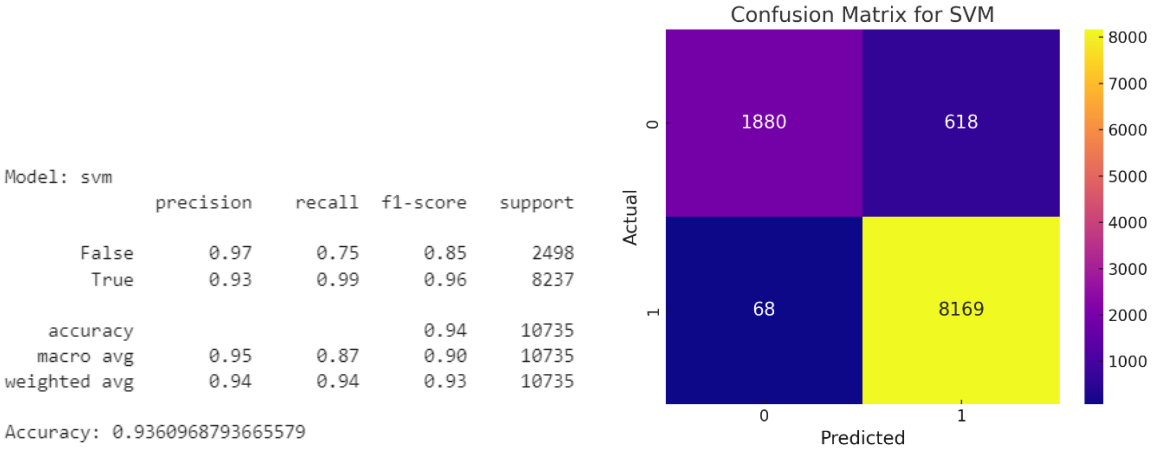


Рисунок 3.6. Результати SVM

Модель випадкового лісу продемонструвала (Рисунок 3.7) високу точність у класифікації з загальною точністю 95.3% на тестовій вибірці,

вказуючи на високу ефективність моделі у розпізнаванні обох класів. Для класу "False" модель показала високу точність (0.98) та досить високий показник відновлення (recall) 0.81, що свідчить про здатність моделі ефективно ідентифікувати негативні випадки з помірною кількістю помилок. Натомість, надзвичайно висока точність (0.95) та майже ідеальне відновлення (1.00) для класу "True" підкреслюють мінімальну кількість помилково негативних результатів, що підтверджується низькою кількістю помилок у матриці помилок.

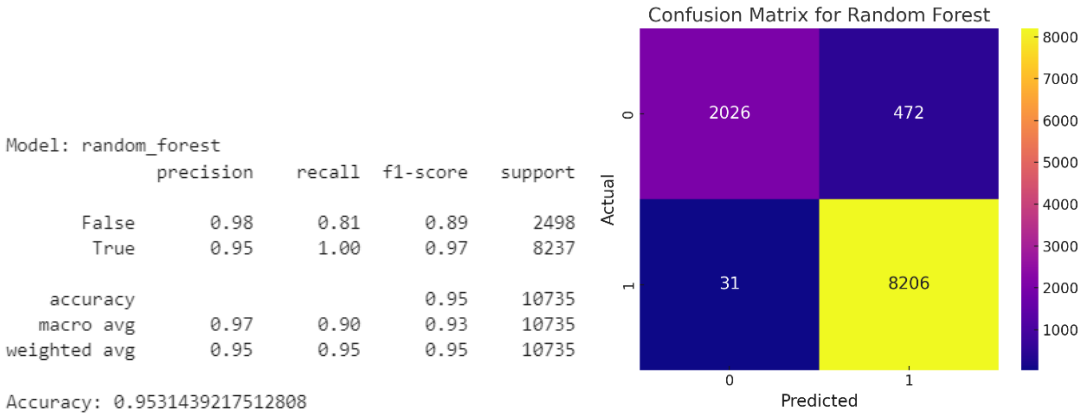


Рисунок 3.7: Результати RandomForestClassifier

Модель градієнтного бустінгу досягла (Рисунок 3.8) загальної точності класифікації 87.3% на тестовій вибірці, що вказує на її здатність відокремлювати класи "True" і "False". Точність моделі для класу "False" була високою (0.94), але показник відновлення (recall) склав лише 0.48, що свідчить про значну кількість помилок другого роду, де негативні випадки часто помилково класифіковані як позитивні. Для класу "True" модель показала вражаючу здатність до класифікації з точністю 0.86 та відновленням 0.99, що вказує на високу ефективність у виявленні істинно позитивних випадків з мінімальною кількістю помилок.

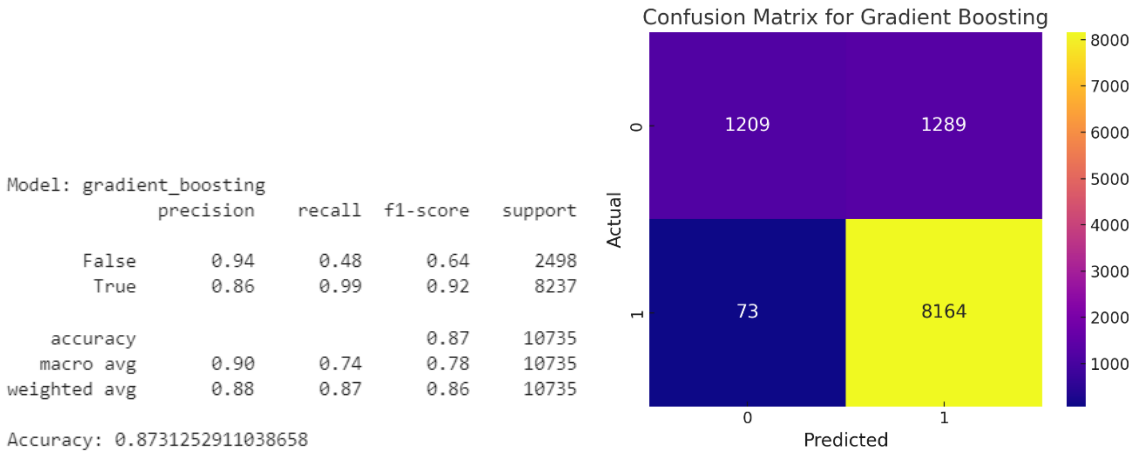


Рисунок 3.8. Результати GradientBoostingClassifier

Модель KNN (метод найближчих сусідів) продемонструвала (Рисунок 3.9) загальну точність класифікації 87.6% на тестовій вибірці, підкреслюючи її спроможність ефективно класифікувати дані. Хоча модель показала високу точність (0.91) для класу "False", показник відновлення (recall) становив лише 0.51, що вказує на складнощі з ідентифікацією всіх негативних випадків. Натомість, для класу "True" модель демонструвала відмінну точність (0.87) та високий показник відновлення (0.99), що свідчить про її здатність з високою вірогідністю ідентифікувати позитивні випадки, хоча і з деякою кількістю помилок першого роду, як це видно з матриці помилок.

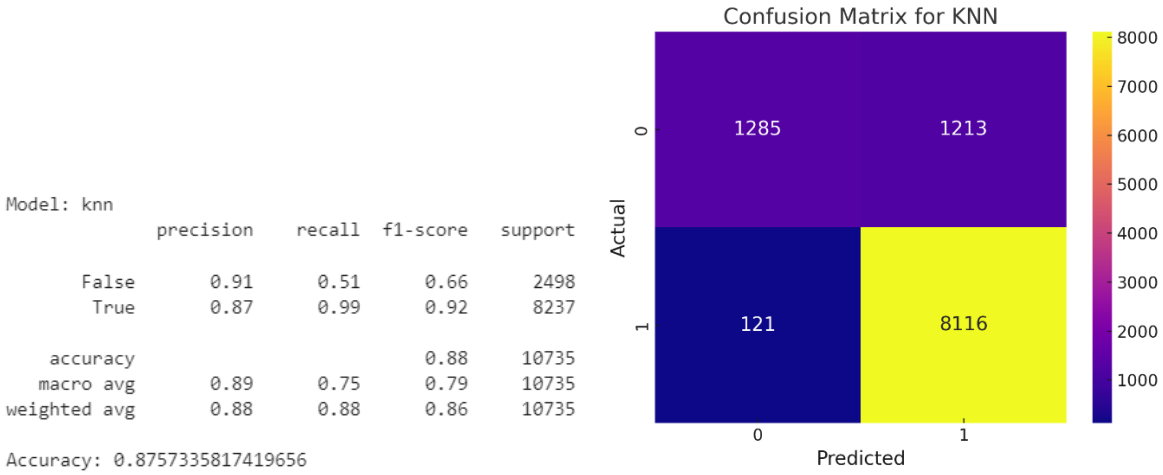


Рисунок 3.9. Результати KNeighborsClassifier

Модель дерева рішень досягла (Рисунок 3.10) високого рівня точності класифікації, 91.4%, на тестовій вибірці, що підтверджує її ефективність у розподілі даних на класи "True" та "False". Для класу "False" модель продемонструвала високі значення точності (0.81) та відновлення (recall) 0.83, що свідчить про її збалансовану здатність виявляти негативні випадки з малою кількістю помилок. Для класу "True", модель досягла вражаючих показників точності (0.95) та відновлення (0.94), що вказує на її надійність у виявленні позитивних випадків з мінімальною кількістю хибнонегативних прогнозів, як це відображено в матриці помилок.

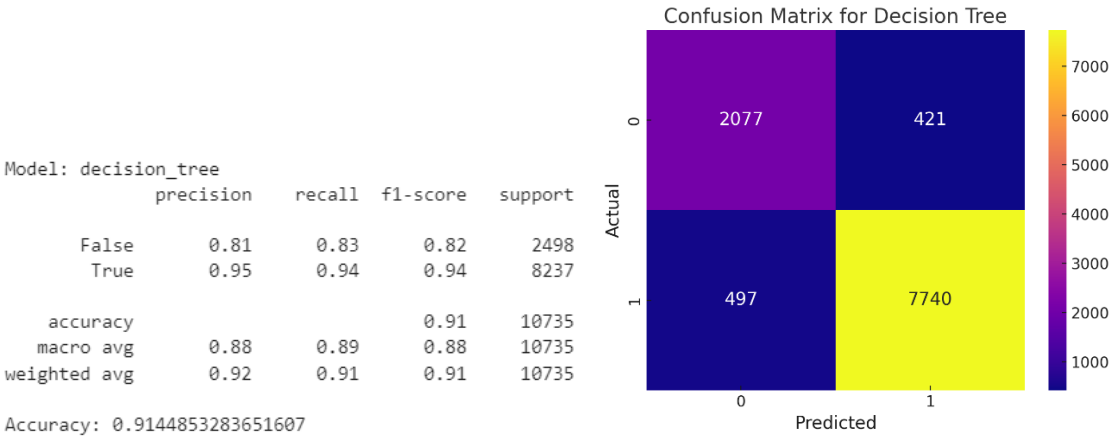


Рисунок 3.10. Результати DecisionTreeClassifier

Модель XGBoost продемонструвала (Рисунок 3.11) загальну точність у 89.2% на тестовій вибірці що вказує на її здатність ефективно класифікувати даний набір даних. Показники точності та відновлення (recall) для класу "False" були 0.89 та 0.61 відповідно, що свідчить про вищу здатність моделі правильно ідентифікувати негативні випадки, хоча і з певною кількістю помилок. Тим часом, для класу "True", модель показала високу точність та відновлення (обидва 0.89 та 0.98), демонструючи відмінну спроможність точно визначати позитивні випадки з мінімальною кількістю помилок, як це відображено в її високому F1-середньому.

Model: xgboost

	precision	recall	f1-score	support
False	0.89	0.61	0.73	2498
True	0.89	0.98	0.93	8237
accuracy			0.89	10735
macro avg	0.89	0.80	0.83	10735
weighted avg	0.89	0.89	0.88	10735

Accuracy: 0.8919422449930136

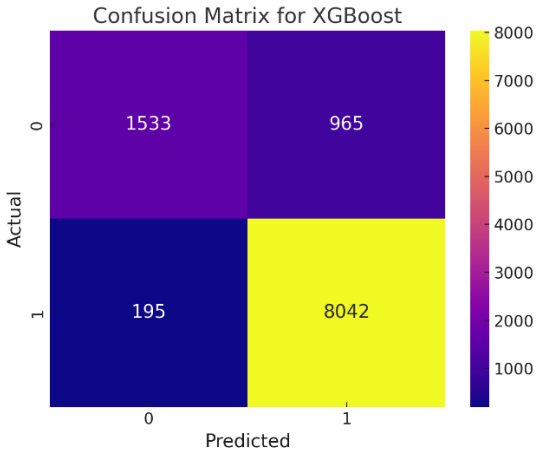


Рисунок 3.11. Результати XGBClassifier

Модель AdaBoost показала (Рисунок 3.12) загальну точність класифікації 86.9% на тестовій вибірці, що свідчить про її ефективність у розпізнаванні даних, хоча і з деякими обмеженнями. Для класу "False" модель забезпечила точність у 0.88 з показником відновлення (recall) у 0.50, вказуючи на відносно низьку здатність до ідентифікації всіх негативних випадків. З іншого боку, висока точність (0.87) та відновлення (0.98) для класу "True" підкреслюють сильну спроможність моделі виявляти позитивні випадки, хоча і з невеликою кількістю помилок, як це відображено в матриці помилок.

Model: adaboost

	precision	recall	f1-score	support
False	0.88	0.50	0.64	2498
True	0.87	0.98	0.92	8237
accuracy			0.87	10735
macro avg	0.88	0.74	0.78	10735
weighted avg	0.87	0.87	0.86	10735

Accuracy: 0.8692128551467163

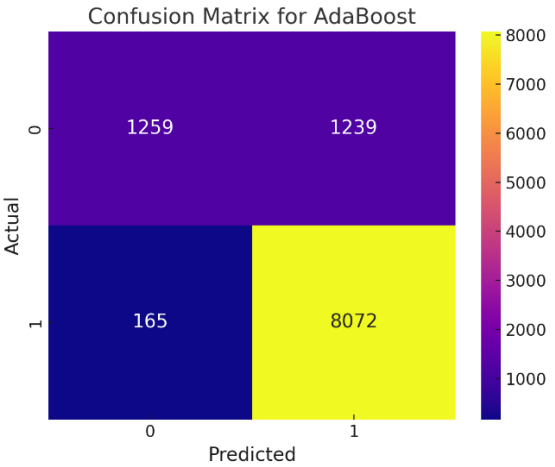


Рисунок 3.12. Результати AdaBoostClassifier

Отже, аналізуючи результати застосування різноманітних моделей машинного навчання до класифікації даних про дезінформацію в українському медіапросторі після початку повномасштабного вторгнення

росії, можна відзначити, що модель випадкового лісу виявилася найбільш ефективною з точністю 95.3%, що підкреслює її здатність високоточно виявляти і розрізняти інциденти дезінформації. Водночас, моделі як-от AdaBoost та логістична регресія показали меншу загальну точність, що може вказувати на їх обмеження у складних умовах ідентифікації тонких випадків дезінформації або більш суб'єктивних аспектів інформаційних операцій. Це підкреслює важливість обрання відповідної моделі для задачі аналізу дезінформації, де випадковий ліс може бути більш придатним для глибокого розуміння та виявлення складних патернів дезінформації в умовах інформаційної війни.

3.4 Висновки з проведеного дослідження

Аналіз ефективності різних моделей машинного навчання, застосованих до задачі класифікації новинних заголовків на предмет дезінформації в українському медіапросторі, демонструє значні варіації в точності, повноті та F1-середньому між моделями. Розглянуті алгоритми, включаючи логістичну регресію, SVM, випадковий ліс, градієнтний бустінг, KNN, дерево рішень, XGBoost та AdaBoost, показали різні рівні продуктивності у вирішенні поставленої задачі.

Модель випадкового лісу виявилася найбільш ефективною, досягнувши загальної точності 95.3%, що вказує на її високу здатність розпізнавати та розрізняти істинні та фальшиві повідомлення. Це підтверджується високими показниками точності для класу "False" (0.98) та класу "True" (0.95), а також значними показниками відновлення для обох класів (0.81 для "False" та 1.00 для "True"), що засвідчує її ефективність у мінімізації як помилок першого, так і другого роду.

Ці результати підкреслюють значення вибору відповідної моделі для конкретної задачі аналізу дезінформації. Вибір моделі не лише впливає на

загальну точність класифікації, але й на здатність моделі мінімізувати помилки першого чи другого роду, що є критично важливим для розробки ефективних інструментів боротьби з дезінформацією. Особливо модель випадкового лісу, яка продемонструвала найкращі результати, може бути рекомендована як оптимальний вибір для подібних задач, забезпечуючи високий рівень точності та здатність ефективно розрізняти між істинними та фальшивими повідомленнями.

Подальші наукові дослідження в області ідентифікації та аналізу дезінформації в українському медіапросторі вимагають глибшого занурення в розробку і вдосконалення алгоритмів машинного навчання, з особливим акцентом на підвищення їх здатності до розпізнавання тонких і складних форм дезінформації. Результати нашого дослідження показують, що модель випадкового лісу з точністю 95.3% виявилася найбільш ефективною, однак існує потенціал для покращення, особливо в аспектах точного розрізнення між істинними та фальшивими повідомленнями. В майбутньому, науковці можуть зосередитись на розробці гібридних моделей, що поєднують переваги кількох алгоритмів, включаючи глибоке навчання та нейронні мережі, для забезпечення більш високої адаптивності та точності в різних інформаційних контекстах. Крім того, важливим напрямком стане розробка методів, що дозволяють моделям краще розуміти семантичний контекст та емоційне забарвлення текстів, що може значно покращити здатність до ідентифікації прихованої дезінформації. Впровадження таких підходів вимагатиме не лише технологічних інновацій, а й більш глибокого розуміння мовних особливостей та культурно-історичних контекстів, на яких базується дезінформація.

Список використаних джерел:

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.

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.

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

14. 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.

15. 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.

16. 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.

17. 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.
18. Gao, B., Zhou, Q., & Deng, Y. (2024). HIE-EDT: Hierarchical interval estimation-based evidential decision tree. Pattern Recognition, 146, 110040.
19. 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.
20. Xing, H. J., Liu, W. T., & Wang, X. Z. (2024). Bounded exponential loss function based AdaBoost ensemble of OCSVMs. Pattern Recognition, 148, 110191.
21. Lipianina-Honcharenko, K., Wolff, C., Sachenko, A., Desyatnyuk, O., Sachenko, S., & Kit, I. (2023). Intelligent Information System for Product Promotion in Internet Market. Applied Sciences, 13(17), 9585.
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. doi:10.47839/ijc.21.2.2596.
22. Gramyak, R., Lipyanina-Goncharenko, H., Sachenko, A., Lendyuk, T., & Zahorodnia, D. (2021). Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring. In IntelITSIS (pp. 246-257).

4

ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ ВИДІЛЕННЯ КЛЮЧОВИХ СЛІВ ДЛЯ КЛАСИФІКАЦІЇ ФЕЙКОВИХ НОВИН

Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Юрків Христина[1], Цюпа Іван[1],Теліховський Олесь[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

4.1 Аналітичний огляд та методологія дослідження фейкових новин

Проблема дезінформації, зокрема поширення фейкових новин, набуває все більшого значення у сучасному інформаційному середовищі. Особливо гостро ця проблема стоїть в умовах гібридних конфліктів та інформаційної війни, де неправдива інформація використовується як інструмент впливу на громадську думку та соціальну стабільність. Тому розробка ефективних методів для автоматичного виявлення фейкових новин є надзвичайно важливою задачею для забезпечення об'єктивного інформування суспільства. У цьому контексті використання машинного навчання та текстових методів для обробки великих обсягів даних відкриває нові можливості для автоматизації процесу ідентифікації неправдивої інформації.

Це дослідження зосереджується на порівнянні різних методів вибору ключових слів для класифікації новин таких, як TF-IDF, RAKE, Yake!, LSA, LDA та TextRank. Основна мета полягає у визначенні найбільш ефективних підходів, які можуть бути використані для створення інструменту

виявлення фейкових новин. За допомогою експериментів із використанням наборів новинних даних було оцінено ефективність кожного методу за такими показниками, як точність (accuracy), відновлення (recall), точність передбачень (precision) та F1-міра. Результати цього дослідження сприятимуть подальшому розвитку технологій автоматичного виявлення дезінформації на основі машинного навчання.

Дана робота зосереджена на порівнянні методів вибору ключових слів для класифікації фейкових і правдивих новин та складається з кількох розділів. У розділі 2 аналізуються існуючі дослідження у сфері виявлення дезінформації та методів обробки тексту. У розділі 3 описано методологію дослідження, включаючи процеси збору даних, попередню обробку та застосування алгоритмів для виділення ключових слів і побудови класифікаційних моделей. У розділі 4 наведено аналіз отриманих результатів з оцінкою точності методів за допомогою метрик precision, recall та F1-міри. Фінальний розділ 5 підкреслює ефективність методів TF-IDF і RAKE та надає рекомендації щодо подальшого розвитку системи для автоматичного виявлення фейкових новин.

Декілька досліджень присвячені порівнянню методів виявлення фейкових новин, зокрема підходів машинного навчання (ML) та глибокого навчання (DL). Наприклад, у дослідженні Халіла Ура та ін. (2023) [1] порівнювали SVM (метод ML) та LSTM (метод DL) для автентифікації новинних статей, де LSTM досяг точності 99,54%, тоді як SVM показав 99,29%. Подібні результати відзначили Чаухан і Палівела (2021) [2], вказавши на високу ефективність нейронних мереж, зокрема LSTM, яка досягла точності 99,88%.

Ці дослідження часто використовують такі підходи, як Random Forest, Decision Tree та методи виділення ознак, як TF-IDF. Наприклад, у

дослідженні Джонсона та ін. (2021)[3] Random Forest досяг 100% точності, перевершивши метод Decision Tree, що досяг 93,64%.

У дослідженні [4] розглянуто три техніки вилучення ознак для виявлення фейкових новин: TF-IDF, Count Vectorizer (CV) та Hash Vectorizer (HV). Ці методи використовують лінгвістичні особливості для покращення ідентифікації дезінформації.

Дослідження [5] пропонує новий підхід до виявлення фейкових новин, заснований на використанні Positive and Unlabeled Learning (PUL) з додаванням механізму уваги для визначення найбільш релевантних ключових слів у мережі новин. Використовується алгоритм GNEE, заснований на Graph Attention Networks (GAT), що дозволяє автоматично вибирати важливі терміни для класифікації новин. Результати показали покращення F1-міри на 2–10% у сценаріях з обмеженою кількістю маркованих даних (тільки 10% фейкових новин були марковані) (Souza et al., 2024).

Більшість порівнянь демонструють переваги методів глибокого навчання, особливо мереж LSTM, які краще обробляють складні текстові патерни завдяки своїй здатності враховувати послідовність даних. Водночас, традиційні моделі ML, як-от SVM, продовжують забезпечувати надійні результати, особливо в умовах обмежених обчислювальних ресурсів. Основною відмінністю даного дослідження є ґрунтовний порівняльний аналіз методів виділення ключових слів таких, як: TF-IDF, RAKE, Yake!, KeyBERT, LSA, LDA та TextRank, у поєднанні з класифікаторами, спрямований на специфічний геополітичний контекст із використанням українських і російських наборів даних, що підвищує його значущість для виявлення фейкових новин під час російсько-української війни.

Враховуючи це, метою даного дослідження є порівняння та визначення найбільш ефективних методів вибору ключових слів для класифікації фейкових та правдивих новин, що дозволить створити точний інструмент для автоматичного виявлення дезінформації. Це дослідження також спрямоване на аналіз різних підходів до обробки тексту таких, як: TF-IDF, RAKE, Yake!, LSA, LDA та TextRank, з метою оцінки їх впливу на точність класифікації та виявлення найбільш результативних для подальшого впровадження у систему виявлення фейкових новин.

4.2 Методологія дослідження

4.2.1 Архітектура дослідження

Рисунок 4.1 ілюструє архітектуру дослідження, яка спрямована на порівняльний аналіз методів обробки тексту для виявлення фейкових новин. На першому етапі відбувається збір даних і їх підготовка, де кожен текстовий документ проходить через процеси токенізації, видалення стоп-слів і приведення тексту до нижнього регістру. Після попередньої обробки дані передаються до блоку виділення ключових слів, який застосовує різні методи для визначення найбільш релевантних термінів у тексті, включаючи TF-IDF, RAKE, Yake!, KeyBERT, LSA, LDA та TextRank. Кожен з цих методів пропонує унікальний підхід до оцінки важливості слів та фраз у документі.

Наступний етап передбачає використання отримання ключових слів для побудови моделей класифікації. Метою цього етапу є оцінка ефективності кожного методу виділення ключових слів у контексті допомоги у виявленні фейкових новин. На завершальному етапі здійснюється оцінка моделі, де вимірюються точність, повнота, F1-міра та інші метрики для кожного методу. Цей підхід дозволяє здійснити всебічний

аналіз різних технік обробки тексту та їх впливу на точність класифікації фейкових новин.

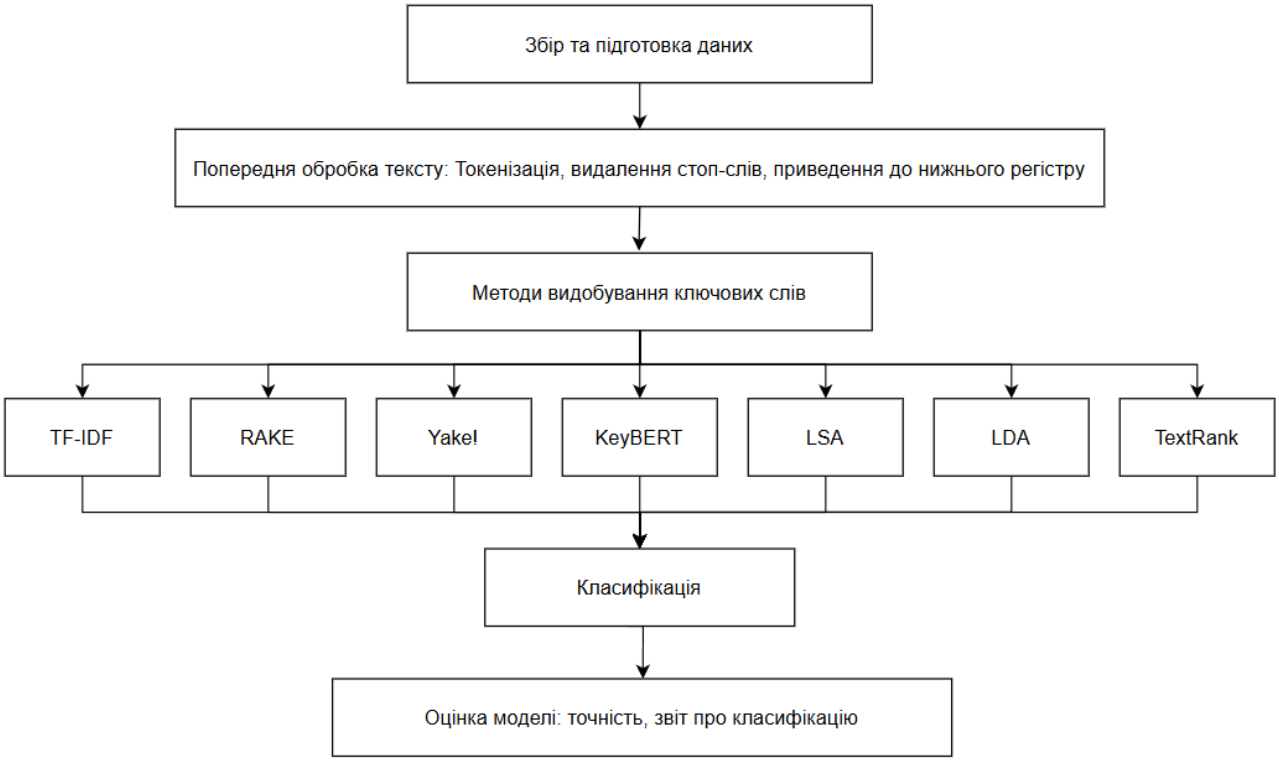


Рисунок 4.1. Структура архітектури дослідження

4.2.2 Опис датасетів

Ukrainian News Fake vs True Dataset і Fake News UA Dataset містять новини з українських та російських джерел, зібрані з метою класифікації фейкових та достовірних новин. Ці набори даних є важливими для аналізу інформаційного простору та досліджень у сфері машинного навчання, що стосуються виявлення фейкових новин під час російсько-української війни.

Ukrainian News Fake vs True Dataset [6] містить приблизно 10 700 заголовків новин, зібраних з українських та російських Telegram-каналів у період з 24 лютого до 11 грудня 2022 року, під час повномасштабного російсько-українського вторгнення. Цей відкритий набір даних включає як достовірні новини, так і фейкові, що публікувалися російськими джерелами. Зокрема, 4 522 заголовків класифіковані як фейкові, а 6 237— як достовірні. Для класифікації новин використовуються два типи міток:

"True" для підтверджених новин і "False" для фейкових. Джерела даних включають такі Telegram-канали, як "Суспільне Новини", "Perepichka NEWS", "NR", а також канали, що поширюють дезінформацію, зокрема "Вокс Україна" та "Война с фейками". Набір даних розроблений для університетського проєкту з метою класифікації новин та є корисним для досліджень у сфері машинного навчання.

Fake News UA Dataset [7] містить зразки новин, зібрані з українських та російських джерел, з метою класифікації фейкових новин. У наборі даних представлено різноманітні статті та заголовки новин з українських Telegram-каналів, а також посилання на першоджерела, такі як Vox Ukraine, Українська правда, Espreso, Радіо Свобода. Кожен запис містить текст новини, її оригінальне джерело, мову (українську або російську), а також мітку, яка вказує на те, чи є новина фейковою. Загалом у наборі близько 12 749 новин, які мають мітки "Fake" (3 375 новин) або "Real". Набір даних призначений для задач класифікації тексту і містить також часові позначки, що дозволяють дослідити динаміку поширення фейкових новин.

Аналіз графіка (Рисунок 4.2) розподілу класів "Fake" та "True" після видалення записів із пропусками показує нерівномірний розподіл новин між фейковими та правдивими у об'єднаному наборі даних. Загалом, кількість правдивих новин (True) становить 4668 (приблизно 58% від загальної кількості), що перевищує кількість фейкових новин (Fake), яких лише 3375 (приблизно 42%).

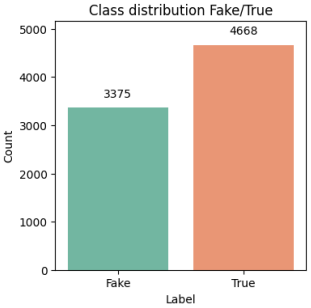


Рисунок 4.2. Розподіл по класах

4.2.3 Опис використаних виділених ключових слів

TF-IDF[8] (Term Frequency-Inverse Document Frequency) — це метод для оцінки важливості терміну в документі стосовно всієї колекції документів або корпусу. Він складається з двох основних частин: TF (частота терміна) та IDF (зворотна частота документа), що разом допомагають виявити, наскільки важливий певний термін у конкретному документі, порівняно з усім набором документів.

Частота терміна показує, як часто певний термін зустрічається в конкретному документі. Формула для розрахунку TF виглядає так:

$$TF(t,d) = \frac{f(t,d)}{\sum_k f(k,d)}$$

де:

$f(t,d)$ — кількість появ терміна t у документі d,

$\sum_k f(k,d)$ — загальна кількість всіх термінів у документі d.

Таким чином, TF показує частку певного терміна від загальної кількості термінів у документі.

Зворотна частота документа вказує на важливість терміна у всьому корпусі документів. Якщо термін зустрічається у багатьох документах, то його важливість знижується. Формула для IDF виглядає так:

$$IDF(t,D) = \frac{\log(|D|)}{(|d \in D : t \in d|)}:$$

$|D|$ — загальна кількість документів у корпусі,

$|\{d \in D : t \in d\}|$ — кількість документів, які містять термін t.

Чим рідше зустрічається термін у документах, тим вищим буде його IDF-значення.

Остаточна формула TF-IDF поєднує обидва ці показники:

$$TF - IDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$$

Це означає, що важливість терміна залежить як від його частоти в певному документі, так і від його загальної поширеності в усьому корпусі документів. Термін, який часто зустрічається в одному документі, але рідко в інших, матиме високе значення TF-IDF. Натомість терміни, що з'являються у багатьох документах, матимуть низьке TF-IDF значення, навіть якщо в конкретному документі вони трапляються часто.

RAKE[9] (Rapid Automatic Keyword Extraction) — це метод для автоматичного вилучення ключових слів з тексту на основі частоти термінів та їх співвідношення з іншими термінами в документі. Він не потребує попереднього навчання на даних або лінгвістичних ресурсів (таких як словники або лексичні бази), що робить його ефективним та швидким для виявлення релевантних ключових слів.

Текст поділяється на фрази, що складаються з одного або декількох слів. Для цього використовуються розділові знаки (.,!?), стоп-слова (загальні слова, які не несуть конкретного смислового навантаження) та інші символи, що не є частиною фраз, які можуть бути ключовими словами. Ці фрази можуть складатися з одного або кількох слів, що стоять поруч.

Для кожного терміна в фразах-кандидатах обчислюються два основні показники:

- Частота терміна f(t): кількість поява терміна у всьому тексті.
- Ступінь терміна d(t): кількість усіх термінів, з якими даний термін знаходиться в одній фразі. Це можна розглядати як суму кількості слів у всіх фразах, що містять цей термін.

Оцінка важливості терміна обчислюється на основі відношення його ступеня до частоти:

$$Score(t) = \frac{f(t)}{d(t)}$$

Це показує відношення між кількістю термінів, що з'являються разом з даним терміном і кількістю його появ у тексті. Чим більший ступінь терміна відносно його частоти в документі, тим важливіше це слово для формування ключової фрази.

Для кожної фрази-кандидата обчислюється її загальна оцінка як сума оцінок всіх термінів, що входять у фразу:

$$Score(Phrase) = \sum_{t \in Phrase} Score(t)$$

Таким чином, фрази, що містять важливі терміни, отримують високі оцінки і розглядаються як потенційні ключові фрази для тексту.

YAKE! (Yet Another Keyword Extractor) [10] — це метод для автоматичного вилучення ключових слів з тексту, який базується на статистичних характеристиках термінів у тексті без необхідності навчання на зовнішніх корпусах або використання лінгвістичних ресурсів. Основна ідея YAKE! полягає в тому, що він оцінює важливість кожного терміна в тексті, застосовуючи кілька показників, які враховують локальний контекст, позицію та частоту терміна.

Спочатку текст розбивається на терміни та фрази-кандидати за допомогою розділових знаків, стоп-слів, та інших символів, що не належать до ключових фраз.

Розраховується частота терміна у тексті, тобто кількість разів, коли термін з'являється в тексті. Чим частіше термін з'являється в документі, тим вища його частотна оцінка:

$$TF(t) = f(t, d)$$

Де $f(t, d)$— кількість появ терміна t у документі d.

YAKE! враховує позицію, на якій з’являється термін у тексті. Термін, що з'являється ближче до початку тексту, може мати іншу вагу, ніж той, що з'являється ближче до кінця.

YAKE! враховує, чи з'являється термін у різних контекстах (фразах). Якщо термін зустрічається лише в обмеженій кількості контекстів, його важливість зменшується.

Оцінюється співвідношення терміна з іншими термінами у фразах-кандидатах. Якщо термін часто з’являється з одними й тими ж іншими термінами, його унікальність зменшується.

Підсумкова формула для оцінки ключової фрази включає зважене врахування частоти терміна, позиції, поширеності та унікальності:

$$Score(t) = W_1 \cdot f(t,d) + W_2 \cdot Pos(t,d) + W_3 \cdot Spread(t)$$

де:

$f(t,d)$ — частота терміна t у документі d,

$Pos(t,d)$— позиція терміна t у документі d,

$Spread(t)$ — кількість контекстів, у яких термін з'являється,

W_1, W_2, W_3 — вагові коефіцієнти для налаштування впливу кожного компонента.

KeyBERT [11] — це метод для вилучення ключових слів із тексту, заснований на використанні моделей трансформерів, зокрема моделі BERT (Bidirectional Encoder Representations from Transformers). На відміну від статистичних методів, таких як TF-IDF або RAKE, KeyBERT використовує семантичне розуміння тексту для ідентифікації ключових фраз і слів, що дозволяє йому враховувати контекст термінів у тексті.

Спочатку текст подається на вхід до моделі BERT, яка генерує багатовимірні векторні представлення (ембедінги) для кожного терміна в тексті. Ці вектори відображають семантичні властивості слів і фраз, що дозволяє оцінювати їхню подібність.

Для вилучення ключових слів KeyBERT використовує векторне подання всього документа або тексту. Потім вектори для кожного слова або фрази в тексті порівнюються з вектором документа для визначення

найбільш релевантних термінів. Якщо вектор фрази або слова найбільш близький до вектора документа, ця фраза або слово вважається потенційним ключовим словом.

Математично, подібність між вектором слова w та вектором документа d розраховується за допомогою косинусної подібності:

$$cosine_{similarity(w,d)} = \frac{w \cdot d}{\| w \| \| d \|}$$

Чим більша косинусна подібність, тим більше слово або фраза відповідає змісту документа.

KeyBERT обчислює косинусну подібність між векторами всіх фраз/термінів і вектором документа. Ті фрази, що мають найвищу подібність, обираються як ключові слова. Це дозволяє не лише оцінювати терміни за їхньою частотою або статистичними показниками, але й враховувати семантичний контекст тексту.

KeyBERT підтримує вилучення багатослівних фраз (n-грам), що дозволяє йому виявляти не тільки окремі слова, але й важливі фрази, що можуть нести більше значення.

LSA [12] (Latent Semantic Analysis) — це метод обробки тексту, який використовується для виявлення прихованих семантичних відношень між термінами в текстових документах. Основною метою LSA є зменшення розмірності простору документів і слів, щоб виявити взаємозв'язки між термінами, що не є очевидними на поверхневому рівні.

Текстовий корпус спочатку перетворюється у матрицю термінів-документів. Кожен рядок матриці відповідає певному терміну, а кожен стовпчик — документу. Значення в комірках можуть бути кількістю появ термінів у документах або зваженими значеннями, наприклад, TF-IDF. Таким чином, початкове представлення тексту є високорозмірним і розрідженим.

Формально, нехай A — матриця термінів-документів розміром m×n,

де m — кількість термінів, а n — кількість документів.

LSA застосовує сингулярний розклад матриці (SVD), щоб зменшити розмірність початкової матриці. SVD розкладає матрицю A на три матриці:

$$A = U\Sigma V^T$$

де:

U — ортогональна матриця термінів,

Σ — діагональна матриця сингулярних значень, що містить ранжовані сингулярні значення (від найбільшого до найменшого),

V^T — ортогональна матриця документів.

Сингулярні значення в матриці Σ визначають важливість кожного з векторів у U і V^T. Менші сингулярні значення можна відкинути, зменшуючи розмірність простору, залишивши лише найзначніші компоненти.

Після виконання SVD, обираються лише найбільші сингулярні значення та відповідні їм компоненти в матрицях U і V^T. Це дозволяє зменшити розмірність матриці, залишивши тільки основні латентні семантичні структури тексту.

Таким чином, матриця зменшеної розмірності дає можливість аналізувати документи і терміни в новому латентному просторі, де терміни, що мають схожі контексти, будуть близькими один до одного.

У зменшеному просторі латентних змінних терміни, які часто з'являються в схожих контекстах, але можуть бути непомітними на поверхневому рівні, стають ближчими один до одного. Це допомагає виявити приховані семантичні зв'язки між термінами та документами, що робить LSA ефективним для задач тематичного моделювання, пошуку інформації та класифікації текстів.

LDA [13] (Latent Dirichlet Allocation) — це статистичний метод тематичного моделювання, який використовується для автоматичного вилучення прихованих тем із великої колекції текстових документів. Основна ідея LDA полягає в тому, що кожен документ розглядається як суміш декількох тем, а кожна тема — як розподіл термінів. Метою алгоритму є виявлення цих тем і розподіл термінів серед них.

Теми в LDA визначаються як розподіли термінів. Це набори слів, які мають високу ймовірність виникнення в межах певної теми. Наприклад, якщо тема стосується політики, то найбільш імовірними словами в ній можуть бути «вибори», «президент», «парламент» тощо.

Кожен документ у корпусі розглядається як суміш кількох тем. Наприклад, документ про економічну політику може бути сумішшю двох тем: економіки та політики.

"LDA припускає, що розподіл тем у кожному документі має заздалегідь визначену форму, наприклад, Dirichlet-розподіл. Цей параметризований розподіл дає змогу контролювати домінування тем у документі за допомогою параметра α, який регулює ймовірність того, наскільки теми рівномірно чи нерівномірно розподіляються в документах.

$$\theta_d \sim Dirichlet(\alpha)$$

де:

θ_d — вектор розподілу тем для документа d,

α — гіперпараметр, що контролює густоту тем у документах.

Кожна тема також має власний розподіл термінів, що теж моделюється за допомогою Dirichlet-розподілу. Параметр β визначає цей розподіл термінів для кожної теми.

$$\phi_k \sim Dirichlet(\beta)$$

де:

ϕ_k — вектор розподілу термінів для теми k,

β — гіперпараметр, що контролює густоту термінів у темах.

Для кожного терміна в документі спочатку обирається тема на основі розподілу тем для цього документа. Потім термін обирається на основі розподілу термінів для обраної теми.

$$z_{d,n} \sim Multinomial(\theta_d)$$

$$w_{d,n} \sim Multinomial\left(\phi_{z_{d,n}}\right)$$

де:

$z_{d,n}$ — тема для n-го терміна в документі d,

$w_{d,n}$ — сам термін, обраний відповідно до теми $z_{d,n}$,

θ_d — розподіл тем для документа d,

$\phi_{z_{d,n}}$ — розподіл термінів для обраної теми $z_{d,n}$.

Основне завдання LDA — знайти розподіл тем для кожного документа і розподіл термінів для кожної теми. Це здійснюється за допомогою методів оцінки параметрів таких, як: Метод очікування-максимізації (EM) або Метод варіаційного висновку.

Генеративна модель LDA описується наступною імовірнісною функцією:

$$P(w, z, \theta, \phi \mid \alpha, \beta) = P(\theta \mid \alpha) \prod_{k=1}^{K} P(\phi_k \mid \beta) \prod_{n=1}^{N} P(z_n \mid \theta) P(w_n \mid z_n, \phi)$$

де:

w — терміни в документах,

z — теми для кожного терміна,

θ— розподіл тем для кожного документа,

ϕ — розподіл термінів для кожної теми,

α, β — гіперпараметри Dirichlet-розподілів.

TextRank [14] — це алгоритм для вилучення ключових слів і автоматичної побудови анотацій тексту, заснований на методах

ранжування графів. TextRank є адаптацією алгоритму PageRank, який використовується в пошукових системах для ранжування веб-сторінок. У випадку TextRank, замість веб-сторінок, алгоритм працює з словами або реченнями тексту.

Текст представлений у вигляді графа, де вершини — це слова або речення, а ребра встановлюються між тими вершинами, які є "близькими" в контексті. Для задач вилучення ключових слів вершинами є окремі слова, для задач сумаризації — цілі речення.

Для вилучення ключових слів:

- Термін вважається вершиною графа.
- Ребро з'єднує два терміни, якщо вони з'являються в межах одного вікна слів (зазвичай вікно має фіксований розмір, наприклад 2 або 3).
- Для кожної вершини v_i, вагу зв'язків з іншими вершинами можна описати через матрицю подібності, яка враховує частоту зустрічі слів разом у вікні.
- Для задач сумаризації:
- Вершинами є цілі речення.
- Ребра з'єднують речення на основі семантичної схожості, яка може бути обчислена, наприклад, за допомогою косинусної подібності між векторними представленнями речень.

Косинусна подібність між реченнями S_i і S_j визначається так:

$$cosine_{similarity(S_i,S_j)} = \frac{S_i \cdot S_j}{\| S_i \| \| S_j \|}$$

Речення, що мають високий ступінь подібності, отримують сильніші зв'язки в графі.

Алгоритм TextRank ітеративно обчислює вагу кожної вершини (слова або речення) за наступною формулою, подібною до формули PageRank:

$$R(v_i) = (1 - d) + d \sum_{v_j \in In(v_i)} \frac{R(v_j)}{(|Out(v_j)|)}$$

де:

$R(v_i)$ — ранг вершини v_i,

d — коефіцієнт загасання (звичайно $d = 0.85$),

$In(v_i)$ — сукупність вершин, які мають зв'язки з v_i,

$Out(v_j)$ — кількість вихідних ребер з вершини v_j,

Ранг кожної вершини розраховується ітеративно до досягнення збіжності.

Ключові слова: Після того як вершини (слова) отримають ранжування, обираються слова з найвищим рангом як ключові.

Сумаризація тексту: Речення з найвищими рангами обираються для побудови анотації тексту.

4.2.4 Класифікатор

На основі попередньої публікації[15] авторів, було проведено порівняльний аналіз різних методів машинного навчання для класифікації фейкових та правдивих новин. Найкращі результати було досягнуто за допомогою класифікатора Random Forest.

Random Forest — це ансамблевий метод, який використовує кілька дерев рішень для покращення продуктивності класифікації. Модель складається з набору дерев рішень $\{T_1, T_2, \dots, T_N\}$, кожне з яких незалежно будується на випадкових підмножинах даних та ознак. Остаточна класифікація здійснюється за допомогою методу голосування (вибір класу, який отримав більшість голосів).

Основні кроки роботи Random Forest:

1. Створення дерев рішень:

- Для кожного дерева випадково вибирається підмножина навчальних даних (із заміною, bootstrapping).
- Вибирається випадкова підмножина ознак для побудови кожного вузла дерева.
- Побудова дерева триває до повної класифікації або до досягнення максимального обмеження глибини.

2. Голосування:

- Для кожного нового прикладу x, кожне дерево T_i в ансамблі прогнозує клас y_i.
- Остаточний прогноз у здійснюється за допомогою більшості голосів:

$$y = \arg\max_{y}(\sum_{i=1}^{N} 1\{T_i(x) = y\})$$

де N — кількість дерев, T_i(x) прогноз дерева T_i для вектора ознак x, а $1\{T_i(x) = y\}$— індикатор того, що дерево T_i прогнозує клас y.

4.3 Результати дослідження

У цьому розділі представлено результати порівняльного аналізу ефективності різних методів класифікації новин за ознакою "Fake" або "True". Досліджувані моделі використовують різні підходи до вибору ключових слів, такі як TF-IDF, RAKE, Yake!, LSA, LDA та TextRank. Метою аналізу було оцінити точність моделей за допомогою метрик precision, recall та f1-score, щоб визначити найкращі алгоритми для виявлення фейкових новин у текстових даних.

Аналіз класифікаційного звіту (Рисунок 4.3.) для моделі на основі TF-IDF показує загальну точність [16] (accuracy) у 0.88 або 88%. Для класу "Fake" модель досягла точності (precision) 0.90, але менший показник відновлення (recall) — 0.81, що свідчить про те, що модель пропускає деякі

фейкові новини. Навпаки, для класу "True" показники точності та повноти значно ближчі: точність становить 0.87, а повнота — 0.94, що означає, що модель добре розпізнає правдиві новини. Підсумкові значення f1-score для класів "Fake" і "True" становлять 0.86 та 0.90 відповідно. Макро-середнє і зважене середнє значення f1-score дорівнює 0.88, що вказує на збалансовану роботу моделі між класами, але з незначною перевагою у виявленні правдивих новин.

```
Accuracy (TF-IDF): 0.8843283582089553
Classification Report (TF-IDF):
              precision    recall  f1-score   support

        Fake       0.90      0.81      0.86       676
        True       0.87      0.94      0.90       932

    accuracy                           0.88      1608
   macro avg       0.89      0.87      0.88      1608
weighted avg       0.89      0.88      0.88      1608
```

Рисунок 4.3. Звіт про класифікацію фейкових та правдивих новин із використанням TF-IDF

Модель класифікації на основі RAKE (Рисунок 4.4) показала загальну точність (accuracy) у 0.88 або 88%. Для класу "Fake" точність становить 0.90, однак повнота трохи нижча — 0.80, що призводить до f1-score 0.85. Клас "True" продемонстрував стабільну точність 0.87 та високу повноту 0.94, що дало результат f1-score 0.90. Макро-середнє і зважене середнє значення f1-score дорівнює 0.87 і 0.88 відповідно, що вказує на стабільну роботу моделі для обох класів. Хоча модель показує збалансовані результати, ефективність у виявленні фейкових новин трохи поступається результатам для правдивих.

```
Accuracy (RAKE): 0.8793532338308457
Classification Report (RAKE):
              precision    recall  f1-score   support

        Fake       0.90      0.80      0.85       676
        True       0.87      0.94      0.90       932

    accuracy                           0.88      1608
   macro avg       0.88      0.87      0.87      1608
weighted avg       0.88      0.88      0.88      1608
```

Рисунок 4.4. Звіт про класифікацію фейкових та правдивих новин із використанням RAKE

Модель Yake! (Рисунок 4.5) продемонструвала точність (accuracy) у 0.78 або 78%, що є нижчим порівняно з іншими моделями. Для класу "Fake" точність склала 0.73, а повнота — 0.76, що дало f1-score на рівні 0.74. Для класу "True" точність була дещо вищою — 0.82, а повнота — 0.79, що призвело до f1-score 0.81. Загальні результати показують, що модель Yake! працює значно гірше, особливо для класу "Fake". Макро-середнє та зважене середнє значення f1-score дорівнює 0.78, що свідчить про менш збалансовану роботу порівняно з іншими методами.

```
Accuracy (Yake!): 0.7804726368159204
Classification Report (Yake!):
              precision    recall  f1-score   support

        Fake       0.73      0.76      0.74       676
        True       0.82      0.79      0.81       932

    accuracy                           0.78      1608
   macro avg       0.77      0.78      0.78      1608
weighted avg       0.78      0.78      0.78      1608
```

Рисунок 4.5. Звіт про класифікацію фейкових та правдивих новин із використанням Yake!

Модель LSA (Рисунок 4.6) продемонструвала точність (accuracy) на рівні 0.85 або 85%. Для класу "Fake" точність становила 0.90, однак повнота

склала 0.72, що призвело до f1-score 0.80. Для класу "True" показники значно вищі: точність — 0.82, повнота — 0.94, і f1-score — 0.88. Хоча загальна точність є високою, відносно низька повнота для фейкових новин свідчить про те, що модель може пропускати деякі фейкові новини. Макро-середнє f1-score становить 0.84, а зважене середнє — 0.85, що свідчить про непогану збалансованість між класами.

```
Accuracy (LSA): 0.849502487562189
Classification Report (LSA):
              precision    recall  f1-score   support

        Fake       0.90      0.72      0.80       676
        True       0.82      0.94      0.88       932

    accuracy                           0.85      1608
   macro avg       0.86      0.83      0.84      1608
weighted avg       0.86      0.85      0.85      1608
```

Рисунок 4.6. Звіт про класифікацію фейкових та правдивих новин із використанням LSA

Модель класифікації на основі LDA (Latent Dirichlet Allocation) показала (Рисунок 4.7) точність (accuracy) на рівні 0.67 або 67%, що є значно нижчим порівняно з іншими моделями. Для класу "Fake" точність становить 0.61, а повнота — 0.56, що дало f1-score 0.59. Для класу "True" точність вища — 0.70, а повнота — 0.74, що забезпечило f1-score на рівні 0.72. Макро-середнє і зважене середнє значення f1-score дорівнюють 0.65 і 0.66 відповідно, що вказує на незбалансованість у роботі моделі, зокрема з гіршим результатом для класу "Fake". Загальні результати показують, що модель LDA має труднощі з ефективною класифікацією фейкових новин.

Модель класифікації на основі TextRank (Рисунок 4.8) показала загальну точність (accuracy) у 0.77 або 77%. Для класу "Fake" точність становить 0.72, а повнота — 0.74, що дає f1-score 0.73. Клас "True" продемонстрував кращі результати: точність — 0.81, повнота — 0.79, і f1-

score — 0.80. Макро-середнє і зважене середнє значення f1-score становлять 0.77, що свідчить про досить збалансовану роботу моделі для обох класів, хоча точність для "Fake" новин нижча, ніж для "True".

```
Accuracy (LDA): 0.6672885572139303
Classification Report (LDA):
              precision    recall  f1-score   support

        Fake       0.61      0.56      0.59       676
        True       0.70      0.74      0.72       932

    accuracy                           0.67      1608
   macro avg       0.66      0.65      0.65      1608
weighted avg       0.66      0.67      0.66      1608
```

Рисунок 4.7. Звіт про класифікацію фейкових та правдивих новин із використанням LDA

```
Accuracy (TextRank): 0.7711442786069652
Classification Report (TextRank):
              precision    recall  f1-score   support

        Fake       0.72      0.74      0.73       676
        True       0.81      0.79      0.80       932

    accuracy                           0.77      1608
   macro avg       0.77      0.77      0.77      1608
weighted avg       0.77      0.77      0.77      1608
```

Рисунок 4.8. Звіт про класифікацію фейкових та правдивих новин із використанням TextRank

Порівнюючи всі методи класифікації за різними метриками, видно, що TF-IDF і RAKE виявилися найефективнішими для класифікації новин, маючи точність 0.8843 та 0.8794 відповідно. Обидва методи продемонстрували збалансовану роботу для класів "Fake" і "True". TF-IDF має високу precision (0.90) для "Fake" новин і дуже високу recall (0.94) для "True" новин, що призводить до загального f1-score 0.88. RAKE, хоч і трохи поступається TF-IDF у повноті для "Fake" новин (recall 0.80), все одно забезпечує стабільну роботу з f1-score 0.85 для "Fake" новин і 0.90 для

"True". Це робить ці два методи найкращими для виявлення фейкових новин у текстових даних.

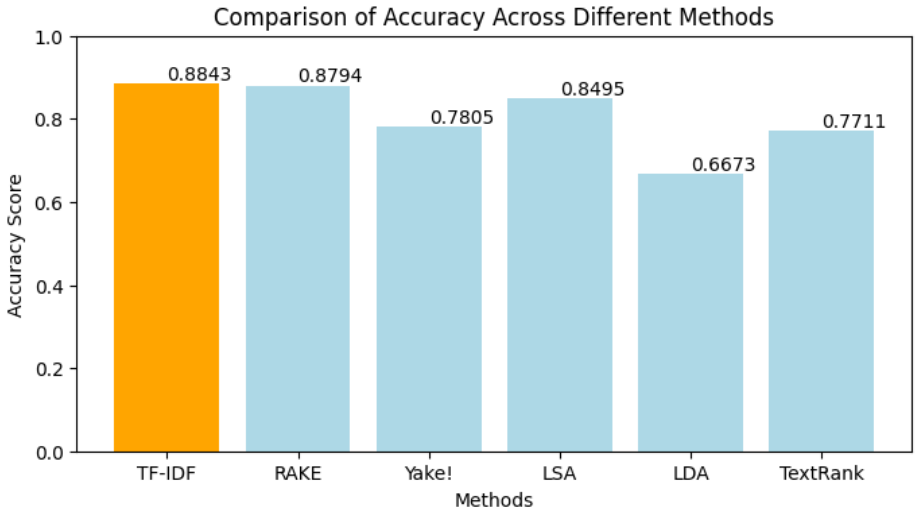


Рисунок 4.9. Порівняльний аналіз точності методів вибору ключових слів для класифікації фейкових та правдивих новин

Результати дослідження показали, що методи TF-IDF та RAKE продемонстрували найвищу ефективність у класифікації новин, маючи точність 0.88. Ці методи показали збалансовану роботу для обох класів ("Fake" і "True"), особливо в плані precision та f1-score. Інші методи, такі як LSA, Yake!, TextRank і LDA, показали гірші результати, зокрема у класифікації фейкових новин, що може бути пов'язано з їхньою меншою здатністю точно визначати ознаки "Fake". Загальний аналіз свідчить про те, що для точнішого виявлення дезінформації підходи TF-IDF та RAKE є найбільш підходящими.

4.4 Висновки з проведеного дослідження

У цьому дослідженні проведено порівняльний аналіз методів вибору ключових слів для класифікації новин як фейкових або правдивих. Використовувалися методи TF-IDF, RAKE, Yake!, LSA, LDA та TextRank,

які застосовувалися до двох наборів новинних даних, що містять заголовки з українських та російських Telegram-каналів. Основна мета дослідження полягала в тому, щоб порівняти ефективність цих методів за допомогою стандартних метрик оцінки класифікації таких, як: precision, recall, f1-score та accuracy. Процес включав попередню обробку тексту, виділення ключових слів, побудову моделей машинного навчання та їх оцінку на основі отриманих результатів.

Кількісні результати показали, що методи TF-IDF та RAKE стали лідерами серед усіх протестованих підходів. TF-IDF продемонстрував найвищу точність — 0.8843, із високою precision для фейкових новин (0.90) та дуже високою recall для правдивих новин (0.94), що забезпечило f1-score 0.88. Метод RAKE мав трохи нижчі показники, із загальною точністю 0.8794 та f1-score 0.85 для фейкових новин і 0.90 для правдивих. Інші методи, такі як LSA (0.8495), Yake! (0.7805), TextRank (0.7711) та LDA (0.6673), показали гірші результати. Особливо низькі результати були отримані при класифікації фейкових новин, що свідчить про труднощі цих підходів в ідентифікації неправдивої інформації.

У подальшому планується розробка інструменту для виявлення фейкових новин, який базуватиметься на результатах цього дослідження. Інструмент буде інтегрувати найефективніші методи вибору ключових слів, зокрема TF-IDF та RAKE, для аналізу текстових даних з новинних джерел. Основною метою буде створення системи, здатної швидко й точно ідентифікувати фейкову інформацію на основі новітніх методів машинного навчання.

Список використаних джерел:

1. Khaleel Ur et al. (2023) "Comparative study of fake news detection between machine learning and deep learning approaches"

https://www.irjmets.com/adminroot/uploaddir/research_conference_paper/document/39/39_2023121236120118.pdf
2. Chauhan and Palivela (2021) "Optimization and improvement of fake news detection using deep learning approaches for societal benefit" https://www.sciencedirect.com/science/article/pii/S2667096821000446?via%3Dihub
3. Johnson et al. (2021) "An Experimental Comparison of Classification Tools for Fake News Detection" https://ijarcce.com/wp-content/uploads/2021/09/IJARCCE.2021.10820.pdf
4. Garg, S., & Sharma, D. K. (2022). Linguistic features based framework for automatic fake news detection. Computers & Industrial Engineering, 172, 108432.
5. de Souza, M. C., Gôlo, M. P. S., Jorge, A. M. G., de Amorim, E. C. F., Campos, R. N. T., Marcacini, R. M., & Rezende, S. O. (2024). Keywords attention for fake news detection using few positive labels. Information Sciences, 663, 120300.
6. https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news
7. https://www.kaggle.com/datasets/sophiamatskovych/fake-news-ua
8. Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45-65.
9. Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. Text mining: applications and theory, 1-20.
10. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, 509, 257-289.
11. Pęzik, P., Mikołajczyk, A., Wawrzyński, A., Nitoń, B., & Ogrodniczuk, M. (2022, November). Keyword extraction from short texts with a text-to-text

transfer transformer. In Asian Conference on Intelligent Information and Database Systems (pp. 530-542). Singapore: Springer Nature Singapore.
12. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.
13. Lipianina-Honcharenko, K., Lendiuk, T., Sachenko, A., Osolinskyi, O., Zahorodnia, D., & Komar, M. (2021, September). An intelligent method for forming the advertising content of higher education institutions based on semantic analysis. In International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications (pp. 169-182). Cham: Springer International Publishing.
14. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
15. Lipianina-Honcharenko, K., Lendiuk, T., Sachenko, A., Osolinskyi, O., Zahorodnia, D., & Komar, M. (2021, September). An intelligent method for forming the advertising content of higher education institutions based on semantic analysis. In International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications (pp. 169-182). Cham: Springer International Publishing.
16. Zhang, M., Li, X., Yue, S., & Yang, L. (2020). An empirical study of TextRank for keyword extraction. IEEE access, 8, 178849-178858.
17. Lipianina-Honcharenko, K., Soia, M., Yurkiv, K., & Ivasechko, A. (2024). Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data.
18. Lipianina-Honcharenko, K., Wolff, C., Sachenko, A., Kit, I., & Zahorodnia, D. (2023). Intelligent Method for Classifying the Level of Anthropogenic Disasters. Big Data and Cognitive Computing, 7(3), 157.

5

МЕТОД ВИЗНАЧЕННЯ НАСТРОЮ ТЕКСТУ ЗА ТЕМАТИЧНИМИ РУБРИКАМИ

Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Мельник Назар[1], Галяс Юрій[1], Лук'янчук Василь[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

5.1 Розробка методу аналізу тональності тексту в контексті інформаційної війни

Актуальність даної статті визначається сучасним контекстом інформаційних воєн, які ведуться на міжнародній арені, з особливим акцентом на війну в Україні. росія використовує інформаційні операції як ключовий елемент своєї гібридної війни проти України та спрямовує значні ресурси на створення та поширення дезінформації, метою якої є вплив на громадську думку, підрив довіри до української держави, її інститутів і лідерів, а також на спотворення реального стану речей у міжнародному співтоваристві.

В контексті посилення інформаційних атак з боку росії, важливість точного та об'єктивного аналізу тональності текстів, пов'язаних з різними аспектами війни та її впливом на різні сфери життя, стає надзвичайно актуальною. Автоматизований аналіз емоційного забарвлення текстів може бути корисним для виявлення спроб маніпулювання громадською думкою, оцінки загального настрою в суспільстві щодо певних подій чи політик, а

також для ідентифікації змін в інформаційному просторі, що можуть вказувати на нові напрямки інформаційних атак чи кампаній дезінформації.

Метод, представлений у статті, інтегрує передові технології обробки природної мови (NLP), машинного навчання та лінгвістичного аналізу, що дозволяє не лише автоматизувати процес визначення емоційного забарвлення текстів, але й забезпечити високу точність і об'єктивність отриманих результатів. Застосування такого комплексного підходу в умовах інформаційної війни відкриває нові можливості для моніторингу інформаційного простору, виявлення та аналізу інформаційних операцій, а також для розробки ефективних стратегій протидії дезінформації.

Враховуючи високу динаміку інформаційного простору та постійну зміну тактик і стратегій інформаційної війни, розробка та впровадження інноваційних методів аналізу текстів, які дозволяють оперативно відстежувати й аналізувати спроби маніпуляції та дезінформації, є надзвичайно важливою. Стаття пропонує такий метод, акцентуючи на його актуальності та необхідності в контексті триваючої війни в Україні та більш ширшого глобального інформаційного протистояння.

Вивчення тональності тексту, яке включає інтеграцію методів обробки природної мови (NLP), машинного навчання та лінгвістичного аналізу, займає важливе місце в сучасних наукових дослідженнях. У [1] розроблено інструмент аналізу тональності тексту, що базується на поєднанні алгоритмів машинного навчання, демонструючи ефективність використання різних методів для точнішого визначення емоційного забарвлення текстів. У [2] підкреслено значення використання TF-IDF та методів машинного навчання для аналізу тональності даних з Twitter (X), що підтверджує важливість цих технік у сфері обробки великих даних. У джерелі [3] розглянуто вплив лексичного багатства навчального корпусу на продуктивність машинного навчання в задачах аналізу тональності,

акцентуючи на необхідності ретельного підходу до вибору та обробки даних. У [4] представлено інноваційний підхід до класифікації ієрархічних зауважень за допомогою BERT та спеціалізованого класифікатора наївного Баєса, що відкриває нові можливості для розширеного аналізу текстів. У [5] запропоновано глибоконавчальний підхід до аналізу тональності для ранжування онлайн-продуктів, використовуючи множини ймовірнісних лінгвістичних термінів, що демонструє потенціал застосування цих технологій у комерційних цілях. У [6] розроблено методику аналізу тональності даних з Twitter (X), використовуючи NLP та машинне навчання, підкреслюючи важливість комплексного підходу до обробки інформації. Автори [7] показують, як аналіз емоцій у Twitter (X) можна вдосконалити за допомогою NLP та методів машинного навчання для категоризації емоцій та візуалізації тенденцій. У [8] розкривають переваги інтеграції лінгвістичного аналізу для дослідження людської мови, що відкриває нові можливості для точної тематичної категоризації. Дослідження, як-от [9], зосереджуються на аналізі онлайн-відгуків про продукти, використовуючи передові технології глибокого навчання та машинного навчання для покращення класифікації даних і вилучення деталізованої інформації про емоції.

Дане дослідження в області аналізу тональності тексту по тематичних рубриках вирізняється глибоким інтегруванням технік обробки природної мови (NLP), машинного навчання та лінгвістичного аналізу, надаючи новизну в порівнянні з аналогами, такими як [10], яка зосереджена на аналізі тональності даних Twitter (X). Відмінною особливістю нашого підходу є розробка спеціалізованих алгоритмів для точного визначення тональності з урахуванням контекстуальної семантики та лексичного різноманіття в межах конкретних тематичних рубрик. Це дозволяє не тільки точніше класифікувати емоційне забарвлення текстів, але й виявляти

тонкі нюанси в емоційних вираженнях, що робить нашу роботу більш адаптованою до складностей природної мови та забезпечує вищу точність аналізу в порівнянні з існуючими методами.

5.2 Метод

Метод визначення тональності тексту за тематичними рубриками є комплексним підходом, що інтегрує обробку природної мови (NLP), машинне навчання та лінгвістичний аналіз для автоматичного класифікування текстових даних. Цей метод дозволяє оцінити емоційне забарвлення тексту (позитивне, нейтральне або негативне) і кількісно виразити його у вигляді відсоткового відношення від -100% до +100%. Далі представимо даний метод наступними етапами та структурно (Рисунок 5.1):

Етап 1. Попередня обробка тексту [11]:

1.1. Видалення зайвих символів, нормалізація тексту;

1.2. Виявлення та видалення стоп-слів;

1.3. Токенізація тексту [11].

Етап 2. Класифікація тексту за тематичними рубриками [12]:

2.1. Використання методів машинного навчання для визначення тематичної рубрики тексту [13];

2.2. Навчання моделей на попередньо визначеному наборі текстів з явною тематичною приналежністю.

Етап 3. Створення та використання словників для кожної тематичної рубрики [14]:

3.1. Розробка словників з ключовими словами та фразами, характерними для кожної рубрики;

3.2. Визначення емоційного забарвлення ключових слів (позитивне, негативне, нейтральне).

Етап 4. Аналіз тональності [15]:

4.1. Використання методів обчислення тональності, таких як оцінювання на основі словників або моделей глибокого навчання;

4.2. Розрахунок індексу тональності T для тексту на основі формули:

$$T = \frac{P - N}{P + N + Q} \times 100\%$$

де P – кількість позитивних слів, N – кількість негативних слів, Q – кількість нейтральних слів.

Етап 5. Інверсія тональності. У випадку, коли текст належить до ворожо налаштованого джерела та не містить прямої згадки про Україну, застосовується інверсія тональності.

Етап 6. Відображення результатів. Розробка інтерфейсу для візуалізації тональності тексту з можливістю перегляду детального аналізу по тематичним рубрикам.

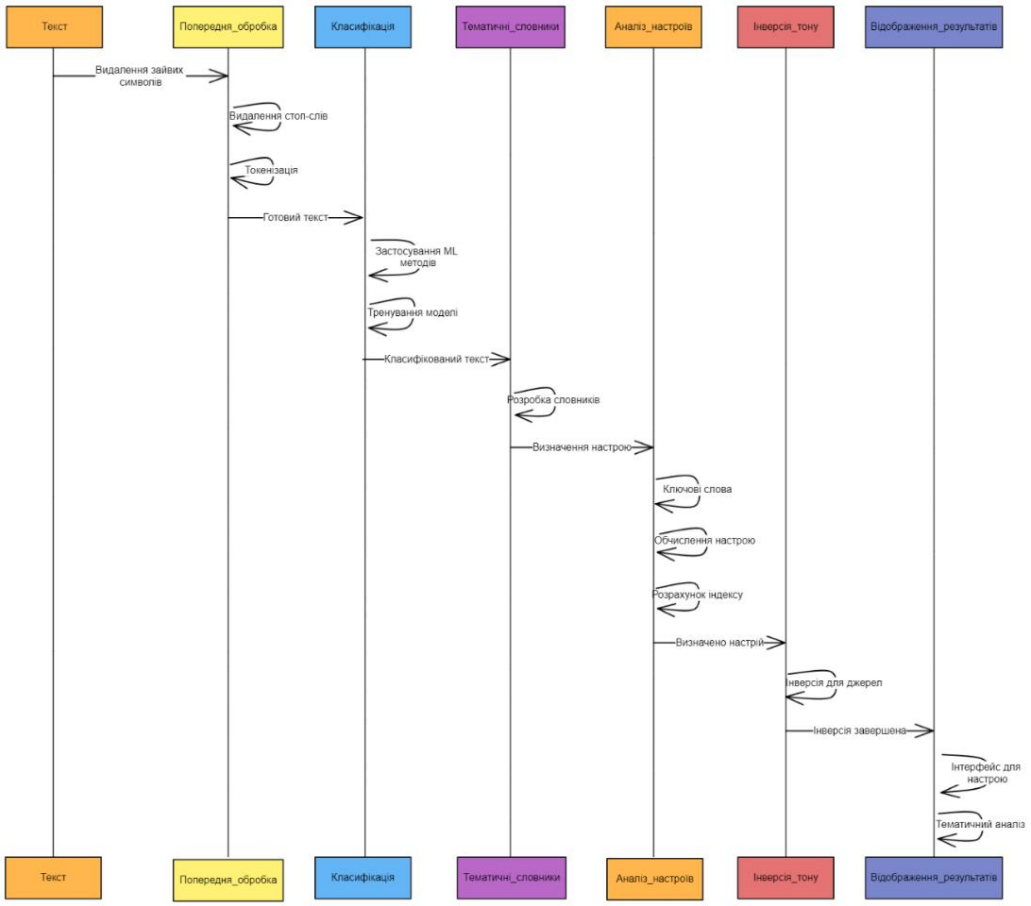


Рисунок 5.1. Структура методу визначення тональності тексту за тематичними рубриками

Далі представлено архітектуру системи аналізу тональності тексту (Рисунок 5.2). Клієнт ініціює взаємодію, надсилаючи запит до сервера, який, у свою чергу, обробляє отриману інформацію. Після завершення обробки сервер пересилає необхідну інформацію назад клієнту. Цей процес є циклічним, тому після передачі інформації клієнт може знову ініціювати взаємодію з сервером. Схема показує типову модель запиту та відповіді, яка є фундаментальною для клієнт-серверної архітектури, де сервер виступає як провайдер ресурсів та послуг, а клієнт — як споживач цих ресурсів та послуг.

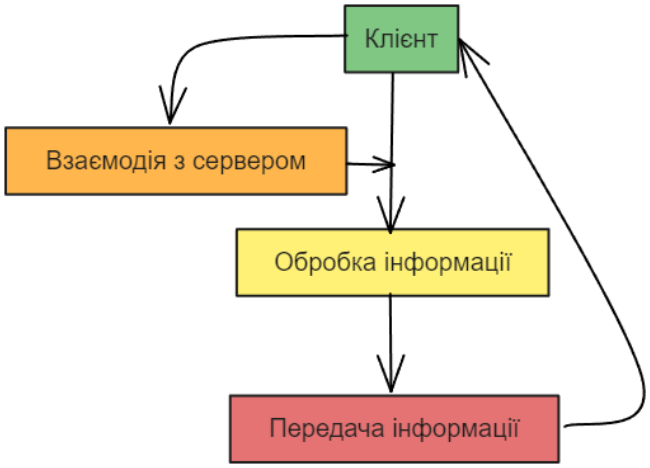


Рисунок 5.2. Клієнт-серверна архітектура класифікації

5.3 Реалізація

Для реалізації запропонованого методу було зібрано набір даних, що охоплює широкого спектру текстів статей з інтернет-платформ України, що забезпечує репрезентативність різних стилів, жанрів та тем. Вибірка охоплює різноманітні емоційні контексти для перевірки здатності алгоритму точно класифікувати емоційні відтінки. Аналіз тексту буде віднесено до однієї з наступних тематичних рубрик, які було визначено експертами з служби безпеки державних органів:

- Військово-політичне керівництво України всіх рівнів;

- Правоохоронні органи України;
- Збройні сили України;
- Соціально-політична ситуація в регіонах України (відношення до мобілізації, соціально-економічна стабільність тощо);
- Проросійські релігійні організації на території України;
- Проросійські рухи, формування концепції «руського миру»;
- Міжнародний імідж України в ЄС (англійська, німецька, польська, румунська, французька, угорська, українська, російська мови);
- Міжнародний імідж України в США, Канаді та Великобританії (англійська мова);
- Міжнародний імідж України в країнах Африки (англійська, французька арабська мови);
- Міжнародний імідж України в країнах Азії (російська, китайська, турецька, арабська, грузинська, казахська, фарсі, киргизька, таджицька, узбецька, японська, корейська мови);
- Україна в інформаційному просторі російської федерації;
- Україна в інформаційному просторі республіки білорусь;
- Соціально-економічна, політична, військова ситуація в російській федерації (відношення до вищого керівництва, мобілізації, погіршення економічного становища тощо).

Основна функціональність веб-інтерфейсу (Рисунок 5.3) [15] дозволяє користувачам вставляти url-посилання сайтів новин для аналізу. Після подачі тексту система використовує алгоритм для визначення емоційного забарвлення, відображаючи результати в формі зрозумілого звіту. Звіт включає кількісні показники присутності позитивних та негативних слів, а також загальний індекс тональності тексту.

Отже, з результату аналізу тональності тексту [16] визначили рубрику "Військово-політичне керівництво України всіх рівнів", де вказано позитивну тональність у +80%. Це відображає високий рівень позитивного емоційного забарвлення тексту статті. Текст статті згадує деталі про військові аспекти, зокрема придбання або використання військової техніки (згадуються F-16 та Bayraktar), та відображає оптимістичне ставлення до подій, пов'язаних з Україною.

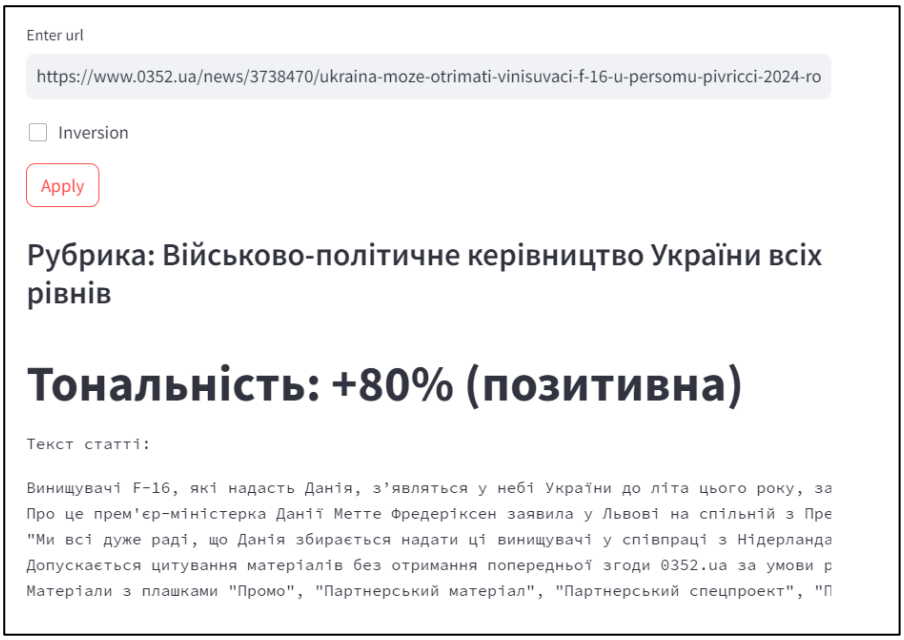


Рисунок 5.3. Приклад результату аналізу тексту

Функція інверсії (Рисунок 5.4) результатів була включена для користувачів, які бажають аналізувати протилежний емоційний відтінок тексту. Це може бути корисним, наприклад, при аналізі текстів, що містять сарказм або іронію, де буквальне значення слів може вводити в оману. Інверсія дозволяє швидко переглянути, як зміниться загальна оцінка емоційного забарвлення, якщо врахувати ці стилістичні фігури.

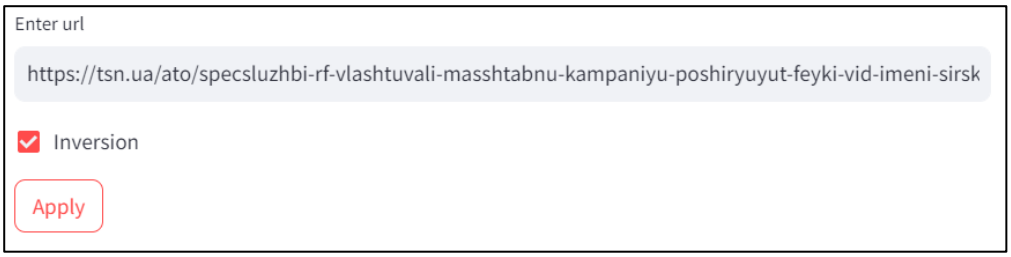
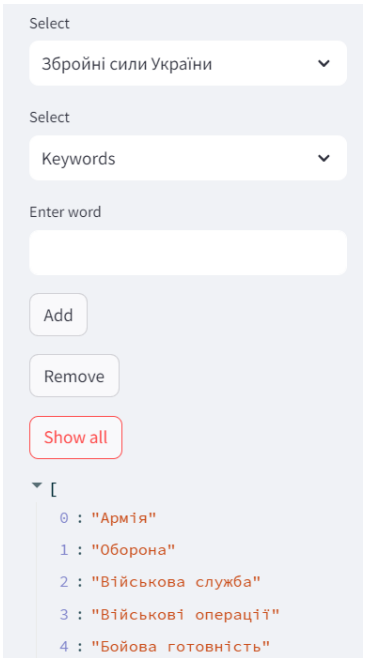


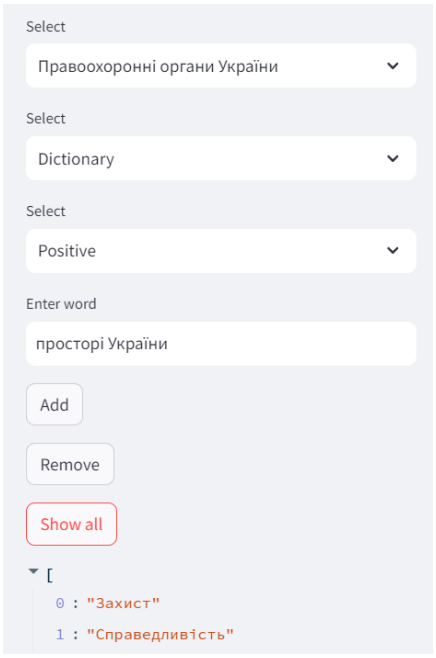
Рисунок 5.4. Поле для введення посилання та використання інверсії

Один з найважливіших аспектів веб-інтерфейсу (Рисунок 5.5а) — можливість редагування ключових слів для кожної рубрики. Користувачі можуть додавати нові слова до словників, вилучати існуючі, що впливає на їх значимість в алгоритмі аналізу. Це дозволяє адаптувати систему до конкретних потреб користувача та підвищити точність визначення тональності для специфічних тем або стилів написання.

Крім того, веб-інтерфейс включає модуль управління словниками (Рисунок 5в), який дозволяє користувачам переглядати та оновлювати базу слів, на основі яких проводиться аналіз. Це особливо важливо для врахування лінгвістичних оновлень, соціальних та культурних змін, які можуть впливати на мову та її емоційне навантаження.



а - ключові слова



в – словник емоцій

Рисунок 5.5. Приклад редагування ключових слів та словника

Було проведено порівняльний аналіз ефективності системи визначення тональності текстів на прикладі 100 новин, шляхом зіставлення автоматизованих результатів системи з оцінками експертів. В аналізі враховувалися такі параметри, як правильність віднесення тексту до відповідної рубрики, порівняння визначеної системою та експертами тональності текстів, а також відповідність цих оцінок. Використання URL як унікального ідентифікатора для кожного тексту забезпечило точне відстеження результатів. Крім кількісних оцінок, експерти надали додаткові примітки для глибшого розуміння причин розбіжностей між експертними оцінками та результатами системи, що сприятиме подальшому вдосконаленню її точності та надійності.

За результатами заповнення таблиці (Таблиця 5.1) обчислено статистичні показники, а саме простий відсоток збігів між системою та експертами.

Таблиця 5.1.

Оцінка ефективності системи

Статистичний показник	Значення
Відповідність рубрик	92%
Середнє відхилення тональності	15%

Результати, представлені у Таблиці 5.1, вказують на досить високу ефективність системи в плані класифікації текстів по рубриках з відповідністю в 92%. Це свідчить про те, що система в більшості випадків правильно ідентифікує тематичну категорію тексту, що підтверджує її надійність у визначенні контексту новин. Однак, середнє відхилення тональності в 15% є досить значним і може вказувати на деякі недоліки у роботі алгоритмів системи з оцінки емоційного забарвлення тексту. Це може бути пов'язано з неповнотою словників, що використовуються для

аналізу тональності. Проте словники можна постійно доповнювати, що є суттєвою перевагою системи, оскільки мова постійно розвивається, а контекст та вживання словникового запасу можуть змінюватися. Можливість доповнення словників користувачами дозволяє системі швидше адаптуватися до нових мовних тенденцій та змін у вживанні термінів, особливо в контексті новин, де важливо враховувати не лише лексичне значення слів, а й їхній конотативний вплив.

Таким чином, система демонструє високу точність у класифікації тематичних рубрик, але потребує поліпшення у визначенні тональності. Постійне доповнення та оновлення словників, з можливістю внесення змін від користувачів, є важливим процесом для підвищення точності аналізу емоційного забарвлення текстів.

5.4 Висновки з проведеного дослідження

У межах цього дослідження було розроблено та реалізовано метод визначення тональності текстів за тематичними рубриками, що поєднує технології обробки природної мови (NLP), машинного навчання та лінгвістичного аналізу. Метод дозволяє не лише класифікувати текст за заданими тематичними категоріями, але й з високою точністю оцінювати його емоційне забарвлення у відсотковому форматі, що є особливо актуальним в умовах сучасної інформаційної війни проти України.

Запропонована система забезпечила високу відповідність класифікації текстів за рубриками — 92%, що підтверджує ефективність моделі в контекстуальній і тематичній ідентифікації. Однак, середнє відхилення у визначенні тональності (15%) свідчить про потребу вдосконалення словників та алгоритмів аналізу емоційного навантаження. Зокрема, інверсія тональності, редагування словників користувачами та

підтримка багатомовності значно підвищують адаптивність та гнучкість системи до нових умов інформаційного простору.

Розроблений інтерфейс користувача забезпечує інтерактивний функціонал для візуалізації результатів аналізу, управління словниками та оперативного реагування на зміну контексту повідомлень. Особливої цінності система набуває при роботі з тематиками, пов'язаними з безпековими викликами, зокрема щодо військово-політичного керівництва, міжнародного іміджу України, проросійських інформаційних кампаній тощо.

Таким чином, результати дослідження підтверджують ефективність запропонованого методу як інструменту інформаційного моніторингу, що здатен підсилити аналітичні можливості у виявленні загроз, пов'язаних із дезінформацією та маніпулятивними впливами. Подальші напрямки розвитку включають удосконалення алгоритмів обробки сарказму, іронії, метафор, розширення мовної підтримки та автоматичне оновлення емоційних словників на основі трендів інформаційного середовища.

Список використаних джерел:

1. Ajitha, P., Sivasangari, A., Rajkumar, R., & Poonguzhali, S. (2021). Design of text sentiment analysis tool using feature extraction based on fusing machine learning algorithms. Journal of Intelligent & Fuzzy Systems, 40(2). https://dx.doi.org/10.3233/jifs-189478

2. Singh, S., Kumar, K., & Kumar, B. (2022). Sentiment Analysis of Twitter Data Using TF-IDF and Machine Learning Techniques. 2022 6th International Conference on Communication and Electronics Systems (ICCES). IEEE. https://dx.doi.org/10.1109/com-it-con54601.2022.9850477

3. Garg, S., Saini, A., & Khanna, N. (2016). Is sentiment analysis an art or a science? Impact of lexical richness in training corpus on machine learning. 2016

International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE. https://dx.doi.org/10.1109/ICACCI.2016.7732474

4. Dhina, M., & Sumathi, S. (2022). An innovative approach to classify hierarchical remarks with multi-class using BERT and customized naïve bayes classifier. International Journal of Engineering, Science and Technology, 13(4). https://dx.doi.org/10.4314/ijest.v13i4.4

5. Liu, Z., Liao, H., Li, M., Yang, Q., & Meng, F. A (2023). Deep Learning-Based Sentiment Analysis Approach for Online Product Ranking With Probabilistic Linguistic Term Sets. IEEE Transactions on Engineering Management. https://dx.doi.org/10.1109/tem.2023.3271597

6. Brindha, K., Senthilkumar, S., Singh, A. K., & Sharma, P. M. (2022). Sentiment Analysis with NLP on Twitter Data. IEEE https://dx.doi.org/10.1109/SMARTGENCON56628.2022.10084036

7. Darshan, K., Samuel, J., Swamy, M., Koparde, P., & Shivashankara, N. (2024). NLP - Powered Sentiment Analysis on the Twitter. https://dx.doi.org/10.36348/sjet.2024.v09i01.001

8. Srinivas, P., Gayathri, K., Bhavitha, K., Jahnavi, & Sarath, K. D. (2023). BLIP-NLP Model for Sentiment Analysis. IEEE https://dx.doi.org/10.1109/ICECAA58104.2023.10212253

9. Bharadwaj, L. (2023). Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification. https://dx.doi.org/10.36948/ijfmr.2023.v05i05.6090

10. Brindha, K., Senthilkumar, S., Singh, A. K., & Sharma, P. M. (2022). Sentiment Analysis with NLP on Twitter Data. IEEE https://dx.doi.org/10.1109/SMARTGENCON56628.2022.10084036

11. Gramyak, R., Lipyanina-Goncharenko, H., Sachenko, A., Lendyuk, T., & Zahorodnia, D. (2021). Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring. In IntelITSIS (pp. 246-257).

12. Lipyanina, H., Sachenko, S., Lendyuk, T., & Sachenko, A. (2020, October). Targeting Model of HEI Video Marketing based on Classification Tree. In ICTERI Workshops (pp. 487-498).
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. Lipianina-Honcharenko, K., Lendiuk, T., Sachenko, A., Osolinskyi, O., Zahorodnia, D., & Komar, M. (2021, September). An Intelligent Method for Forming the Advertising Content of Higher Education Institutions Based on Semantic Analysis. In International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications (pp. 169-182). Cham: Springer International Publishing.
15. Наша розробка https://nelczgkwcsghtwdkw7buxn.streamlit.app/
16. https://www.0352.ua/news/3738470/ukraina-moze-otrimati-vinisuvaci-f-16-u-persomu-pivricci-2024-roku-premerka-danii

6

АНСАМБЛЬ КЛАСИФІКАТОРІВ ДЛЯ ОНЛАЙН ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ

Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Івасечко Андрій[1], Кошицький Костянтин[1], Лендюк Дмитро[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

6.1 Аналіз попередніх досліджень

У період російсько-української війни особливо важливим стає виявлення фейкових новин, оскільки дезінформація може призвести до серйозних наслідків. Дослідження Тао В. та Пенг Й. (2023) разом з роботою Майнича С., Булгакової А. та Висоцької В. (2023) підкреслюють роль соціальних мереж, таких як Twitter (X) та Weibo, у формуванні громадської думки під час конфліктів. Вони вказують на інтенсивне зростання активності на початку війни, яке хоч і знизилося, все ще залишається значущим для сприйняття України як жертви агресії. Відтак, стає життєво необхідним відрізняти реальні новини від фейкових, особливо коли соціальні медіа слугують основним джерелом інформації для багатьох людей. Це підвищує потребу у розробці дієвих методів виявлення фейкових новин, що дозволить забезпечити достовірність інформації та загальну безпеку. Дослідження пропонує інноваційний метод, який об'єднує переваги класифікаторів у ансамблях з техніками онлайн навчання, забезпечуючи точність ідентифікації фейкових новин та адаптацію до

динамічно змінюваного інформаційного середовища, що дозволяє проводити ефективне виявлення дезінформації в реальному часі.

В цьому дослідженні представлена розробка та аналіз новаторського методу ансамблю класифікаторів для онлайн виявлення дезінформації (ECOD), що об'єднує переваги адаптивного онлайн навчання з динамічним ковзним вікном для ансамблів класифікаторів тексту. Метою є підвищення точності та адаптивності в ідентифікації фейкових новин, зокрема в українськомовному інформаційному просторі. Ця робота акцентує увагу на важливості розробки інструментів, здатних адекватно працювати з унікальними лінгвістичними та культурними особливостями конкретних мовних груп, а також на необхідності забезпечення високої реактивності систем у відповідь на стрімкі зміни в інформаційному потоці.

Vasist, P.N., та Krishnan, S. (2023)[3] досліджують зв'язок між фейковими новинами та інноваціями, орієнтованими на сталість, вказуючи на необхідність розробки цілісної концептуальної рамки для розуміння цього взаємозв'язку. Автори визначають ключові теоретичні та практичні наслідки, а також пропонують напрямки майбутніх досліджень у цій галузі.

Hamed, S. K., Ab Aziz, M. J., та Yaakub, M. R. (2023)[4] та Kondamudia, M. R., Sahoob, S. R., Chouhanc, L., та Yadavd, N. (2023)[5] проводять оглядові дослідження, аналізуючи різні підходи до виявлення фейкових новин. Hamed та співавтори зосереджуються на викликах, пов'язаних з наборами даних, представленням ознак та інтеграцією даних, тоді як Kondamudia та співавтори розглядають атрибути, особливості та методи виявлення фейкових новин у соціальних мережах, включаючи лінгвістичний та семантичний аналіз.

У дослідженнях Hu, L., Wei, S., Zhao, Z., та Wu, B. (2022)[6] та Phan, H. T., Nguyen, N. T., та Hwang, D. (2023)[7] акцентується на використанні глибокого навчання та графових нейронних мереж для виявлення фейкових

новин. Hu та співавтори розглядають різні підходи глибокого навчання, включаючи кероване, слабо кероване та некероване навчання, та аналізують їх ефективність на основі різних наборів даних. Phan та співавтори зосереджуються на використанні графових нейронних мереж, вказуючи на їх потенціал та виклики, пов'язані зі стандартизацією даних та розробкою спеціалізованого апаратного забезпечення.

Дослідження Das, B., та TSB, S. (2023)[8] висвітлюють важливість багатоконтекстуального навчання у дослідженні дезінформації, враховуючи різні контексти, такі як зміст, емоції, користувачі та інші. Автори пропонують всебічний погляд на цю проблему, включаючи виклики, пов'язані з багатомодальністю контенту та недостатністю мічених даних. Ruffo, G., Semeraro, A., Giachanou, A., та Rosso, P. (2023)[9] також підкреслюють необхідність інтердисциплінарного підходу, включаючи аналіз мереж та мови, для розуміння динаміки поширення фейкових новин та їх впливу.

У статті Baker, M., Jihad, K., & Taher, Y. (2023)[10] досліджується, як люди висловлюють свої почуття на платформі Twitter (X) під час конфлікту між росією та Україною. Дослідження має дві цілі: збір унікальних даних та використання машинного навчання (ML) для класифікації твітів залежно від їх впливу на почуття людей. Перша ціль полягала у виявленні найбільш релевантних хештегів, пов'язаних з конфліктом, для знаходження датасету. Друга ціль - використання декількох відомих моделей ML для групування твітів. Експериментальні результати показали, що більшість класифікаторів ML мають вищу точність на збалансованому наборі даних. Однак результати експериментів з використанням стратегій балансування даних не обов'язково вказують на те, що всі класи будуть працювати краще. Тому важливо підкреслити важливість порівняння та контрастування стратегій балансування даних, використаних у дослідженнях SA (Sentiment

Analysis) та ML, включаючи більше класифікаторів та ширший спектр випадків використання.

У статті автори Chang, Q., Li, X., та Duan, Z. (2024)[11] пропонують новий підхід до виявлення фейкових новин за допомогою глибокого навчання, використовуючи техніки обробки природної мови (NLP) для кодування вузлів з контекстом новин та користувачів. Вони застосовують триграфові згорткові мережі для вилучення інформативних ознак з мережі поширення новин та агрегують внутрішню та зовнішню інформацію користувачів. Методика включає глобальний механізм уваги з пам'яттю для вивчення структурної однорідності мереж поширення новин, а також модуль для агрегації часткової ключової інформації. Експериментальні результати показують ефективність запропонованого підходу у виявленні фейкових новин, досягаючи високих показників точності та F1-оцінки на реальних наборах даних. Наприклад, модель GCN-GANM показала високу точність (Acc.) та F1-оцінку на датасетах Gossipcop (Acc.: 0.9825, F1: 0.9825) та Weibo (Acc.: 0.9804, F1: 0.9805), перевершуючи інші методи. Ця робота пропонує новий напрямок у дослідженні виявлення новин, поєднуючи глобальну та часткову інформацію.

У статті автори Peng, L., Jian, S., Kan, Z., Qiao, L., та Li, D. (2024)[12] розробили новий метод виявлення фейкових новин, який враховує контекстуальну семантичну репрезентацію для аналізу багатомодальних даних у соціальних мережах. Цей метод, названий CSFND (Contextual Semantic representation learning for multimodal Fake News Detection), включає некерований етап навчання контексту для отримання локальних контекстуальних ознак новин, які потім поєднуються з глобальними семантичними ознаками для навчання контекстуальної семантичної репрезентації новин. Експерименти на двох реальних багатомодальних наборах даних показали, що CSFND значно перевершує десять сучасних

методів виявлення фейкових новин, покращуючи середню точність на 2,5% порівняно з найкращими базовими методами.

У статті автор Ahammad, T. (2024)[13] досліджує розповсюдження фейкових новин про COVID-19, яке стало критичною проблемою. Дослідження спрямоване на отримання уявлення про типи поширюваної дезінформації та розробку глибокого аналітичного підходу для аналізу фейкових новин про COVID-19. Воно поєднує аналіз сентиментів (SA) та моделювання тем (TM) для покращення точності вилучення тем з великого обсягу неструктурованих текстів, враховуючи сентимент слів. Було зібрано та підготовлено набір даних, що містить 10,254 заголовків новин з різних джерел, і застосовано правило SA для маркування набору даних трьома сентимент-тегами. Серед оцінених моделей TM, Latent Dirichlet Allocation (LDA) продемонструвала найвищий показник згуртованості 0.66 для 20 згуртованих тем з негативним сентиментом та 0.573 для 18 згуртованих позитивних фейкових новин, перевершуючи Non-negative Matrix Factorization (NMF) (згуртованість: 0.43) та Latent Semantic Analysis (LSA) (згуртованість: 0.40). Виявлені теми підкреслюють, що дезінформація переважно обертається навколо вакцини від COVID, злочинів, карантину, медицини та політичних та соціальних аспектів. Це дослідження пропонує уявлення про вплив фейкових новин про COVID-19, надає цінний метод для виявлення та аналізу дезінформації та підкреслює важливість розуміння шаблонів та тем фейкових новин для захисту громадського здоров'я та сприяння науковій точності.

У статті автори Qu, Z., Meng, Y., Muhammad, G., та Tiwari, P. (2023)[14] розробили нову модель для виявлення фейкових новин на соціальних мережах, засновану на квантовому багатомодальному злитті (QMFND). Модель QMFND інтегрує вилучені зображення та текстові ознаки, які проходять через запропоновану квантову згорткову нейронну

мережу (QCNN), щоб отримати дискримінативні результати. Перевірка QMFND на двох наборах даних соціальних медіа, Gossip та Politifact, показала, що її продуктивність дорівнює або навіть перевищує продуктивність класичних моделей. Крім того, було досліджено вплив різних параметрів. QCNN виявилася не тільки високовиразною та здатною до заплутування, але й стійкою до квантового шуму.

У статті автори Farhangian, F., Cruz, R. M. O., та Cavalcanti, G. D. C. (2024)[15] провели детальне дослідження в області виявлення фейкових новин, запропонувавши оновлену таксономію для цієї галузі, засновану на кількох критеріях: типах використаних ознак, перспективах виявлення фейкових новин, методах представлення ознак та підходах до класифікації. Автори провели широкомасштабне емпіричне дослідження, оцінюючи різні техніки представлення ознак та підходи до класифікації на основі точності та обчислювальних витрат. Результати експериментів показали, що оптимальні техніки вилучення ознак залежать від характеристик набору даних. Моделі, засновані на трансформерах, послідовно демонстрували вищу продуктивність. Крім того, використання трансформерних моделей, як методів вилучення ознак, а не просто тонкого налаштування мережі для післядіяльності, покращує загальну продуктивність. Ретельний аналіз помилок показав, що комбінація методів представлення ознак та алгоритмів класифікації, включаючи класичні, пропонує доповнюючі аспекти, які повинні бути враховані для досягнення кращої загальної продуктивності при збереженні низьких обчислювальних витрат.

У статті Fang, X., Wu, H., Jing, J., Meng, Y., Yu, B., Yu, H., & Zhang, H. (2024)[16] представлено новий підхід до раннього виявлення фейкових новин через сприйняття семантичного середовища новин (NSEP). NSEP використовує графові згорткові мережі для виявлення семантичних невідповідностей між вмістом новин та зовнішніми постами, а також

модуль мікросемантичного виявлення з багатоголовою та розрідженою увагою для виявлення семантичних суперечностей. Експерименти на реальних китайських та англійських датасетах показали, що NSEP досягає точності до 86.8% на китайських датасетах, що на 14.1% вище, ніж у інших методів. Це підтверджує ефективність виявлення фейкових новин через аналіз мікро- та макросемантичних середовищ.

У статті Soga, K., Yoshida, S., & Muneyasu, M. (2023)[17] розглядається проблема виявлення фейкових новин на соціальних мережах. Автори пропонують новий підхід, який враховує схожість думок користувачів, аналізуючи їхні позиції щодо новинних статей та взаємодії в постах. Використовуючи мережу графових трансформерів, метод одночасно вилучає глобальну структурну інформацію та взаємодії схожих позицій. Метод був оцінений на спеціально зібраних даних з Twitter (X) та на датасеті FibVID, демонструючи значне покращення у порівнянні з традиційними методами, включаючи найсучасніші методи.

У статті Yang, H., Zhang, J., Zhang, L., Cheng, X., & Hu, Z. (2024)[18] розглядається проблема виявлення фейкових новин у мультимодальних даних. Автори вказують на виклики, пов'язані з аналізом взаємозв'язків між регіонами зображень та фрагментами тексту, а також на необхідність більш глибокого аналізу ієрархічної семантики тексту. Для вирішення цих проблем пропонується мультимодальна мережа з урахуванням взаємозв'язків (MRAN), яка включає в себе кілька етапів обробки. Спочатку використовується багаторівнева мережа кодування для вилучення семантичних особливостей тексту, а також VGG19 для вилучення візуальних характеристик. Потім використовується мережа уваги, яка обчислює схожість між сегментами інформації в межах модальностей та між модальностями. Нарешті, отримані характеристики передаються в детектор фейкових новин. Експерименти на трьох датасетах показали

ефективність MRAN, підкреслюючи його сильну продуктивність у виявленні фейкових новин.

У статті Raja, E., Soni, B., Lalrempuii, C., & Borgohain, S. K. (2024)[19] розглядається проблема виявлення фейкових новин у мовах дравідійської групи, які є малоресурсними. Автори пропонують гібридну модель глибокого навчання, що інтегрує розширені тимчасові конволюційні нейронні мережі (DTCN), двонаправлені довготривалі короткотермінові пам'яті (BiLSTM) та механізм уваги з контекстуалізацією (CAM). DTCN використовується для вловлювання тимчасових залежностей, BiLSTM - для ефективного захоплення довготривалих залежностей, а CAM - для акцентування важливої інформації та зменшення впливу нерелевантного контенту. Також в моделі використовується адаптивний циклічний навчальний коефіцієнт з механізмом ранньої зупинки для покращення збіжності моделі. Результати показують, що запропонована модель перевершує існуючі методи та досягає високої середньої точності 93.97% на датасеті Dravidian_Fake у чотирьох дравідійських мовах.

У статті Luvembe, A. M., Li, W., Li, S., Liu, F., & Wu, X. (2024)[20] розглядається проблема виявлення фейкових новин у мультимодальному контексті. Автори вказують на недоліки існуючих методів, які інтегрують крос-модальні особливості, не враховуючи некорельовані семантичні представлення, що може додавати шум до мультимодальних характеристик. Це знижує точність моделей, оскільки важливо враховувати тонкі відмінності між текстом та зображеннями для ідентифікації фейкових новин. Для вирішення цих викликів запропоновано модель CAF-ODNN, яка використовує доповнювальну увагу та оптимізовану глибоку нейронну мережу для виявлення тонких крос-модальних зв'язків. Модель включає генерацію підписів до зображень для створення семантичних представлень, двонаправлену доповнювальну увагу між модальностями та компонент

вирівнювання і нормалізації для калібрування об'єднаних представлень. Оптимізована глибока нейронна мережа (ODNN) використовується для покращення екстракції ознак. Модель перевершує аналогічні підходи на стандартних метриках на чотирьох реальних наборах даних, підкреслюючи важливість доповнювальної уваги та оптимізації у виявленні фейкових новин.

Стаття Jiang, Y., Yu, X., Wang, Y., Xu, X., Song, X., & Maynard, D. (2023)[21] розглядає новий підхід до виявлення фейкових новин, який використовує багатомодальне навчання з використанням шаблонів запитів (prompts). Традиційно, виявлення фейкових новин здійснювалося за допомогою аналізу текстової інформації, але цей підхід часто не враховує складність та тонкощі онлайн-дезінформації. Новий підхід, запропонований у статті, включає використання багатомодальних даних (текст та зображення) та методику навчання з використанням шаблонів запитів, що дозволяє краще виявляти фейкові новини, особливо в умовах обмежених даних. Автори використовують три типи шаблонів запитів з м'яким вербалізатором та метод злиття, що враховує схожість, для адаптивного об'єднання багатомодальних представлень. Цей підхід демонструє вищі показники F1 та точності на двох багатомодальних стандартних наборах даних, підтверджуючи його ефективність у реальних сценаріях.

Стаття Syed, L., Alsaeedi, A., Alhuri, L. A., & Aljohani, H. R. (2023)[22] пропонує гібридний підхід, який поєднує слабо наглядоване навчання та глибоке навчання для виявлення фейкових новин у соціальних мережах. Дослідження зосереджується на використанні методів машинного навчання, таких як SVM, для анотації великих обсягів нерозмічених даних, а також на застосуванні глибоких навчальних технік, таких як Bi-LSTM та Bi-GRU, для класифікації фейкових новин. Автори використовують TF-IDF

та Count Vectorizer для вилучення ознак з текстових даних. Експериментальні результати показують, що запропонований підхід досягає високої точності у виявленні фейкових новин, що робить його ефективним інструментом у боротьбі з дезінформацією в інтернеті.

Стаття Xie, B., & Li, Q. (2023)[23] вводить концепцію "воротарів" у соціальних мережах для виявлення фейкових новин. Воротарі – це активні користувачі, які беруть участь у поширенні новин. Дослідження пропонує модель поведінки воротарів на основі рекурентної нейронної мережі (RNN), яка включає навчання моделі та виявлення фейкових новин. Метод здатний виявляти фейкові новини в реальному часі, використовуючи дані з Twitter (X) та Weibo. Експериментальні результати показують, що одиниця повторюваного воротаря (GRU) досягає найкращого загального результату. Запропонований метод перевершує декілька сучасних підходів, демонструючи ефективність у ранніх та середніх стадіях поширення новин.

У статті Пшібаня, Герціга та Штайнбергера (2019)[24] розглядається задача автоматизації виявлення фейкових новин та перевірки фактів для західнослов'янських мов, зокрема для чеської, польської та словацької. Автори представляють набори даних для цих мов та проводять початкові експерименти, які встановлюють базовий рівень для подальших досліджень у цій галузі. Вони використовують 10-кратну крос-валідацію для оцінки як збалансованих, так і незбалансованих наборів даних, а також проводять бінарні експерименти лише з класами "ПРАВДА" та "НЕПРАВДА". В якості вхідних даних для класифікатора використовуються текст твердження або текст твердження, доповнений текстом обґрунтування.

У статті Bucos, M., & Drăgulescu, B. (2023)[25] досліджується ефективність використання методу зворотного перекладу (Back Translation, BT) з використанням трансформерних моделей для покращення виявлення фейкових новин румунською мовою. Дослідження базується на даних з

Factual.ro, де моделі з BT показали кращі результати за точністю, прецизійністю, відгуком, F1-оцінкою та AUC порівняно з моделями, навченими на оригінальному наборі даних. Використання mBART для BT з французькою як цільовою мовою покращило продуктивність моделі порівняно з Google Translate. Найкраще серед тестованих моделей показали себе Класифікатор Екстра Дерев та Класифікатор Випадкових Лісів. Результати свідчать про потенціал використання BT з трансформерними моделями, такими як mBART, для підвищення ефективності виявлення фейкових новин.

У статті Afanasieva, I., Golian, N., Golian, V., Khovrat, A., & Onyshchenko, K. (2023).[26] досліджується проблема визначення достовірності інформації, особливо в контексті соціальних заворушень та важливих подій, таких як вибори президента США та вторгнення росії в Україну. Автори розглядають ефективність використання нейронних мереж для виявлення фейкових новин. Для підвищення точності класифікації було створено алгоритм попередньої обробки даних на основі основних принципів обробки природної мови. В результаті дослідження були виявлені мовні шаблони фейкових новин, які стали основою для попередньої обробки даних. Описано особливості конволюційних та рекурентних нейронних мереж та їх модифікації для аналізу текстових даних. Для порівняння певних моделей було обрано набір показників, що характеризують ефективність алгоритмів. Точність класифікації цих моделей була перевірена на даних, пов'язаних з виборами президента США та масштабним вторгненням російської федерації на територію України.

На основі аналізу сучасних досліджень у галузі виявлення дезінформації, представлених вище, можна зробити висновок, що існуючі підходи значною мірою концентруються на англійській та китайській мовах, залишаючи українську мову з недостатнім рівнем уваги. Це вказує

на відсутність спеціалізованих наборів даних та моделей, які були б адаптовані до особливостей української мови. У зв’язку з цим, розроблено підхід "Ансамбль класифікаторів для онлайн виявлення дезінформації" (ECOD), який спрямований на створення комплексного методу для виявлення дезінформації в україномовному контенті. Метод включає етапи збору та попередньої обробки даних, аналізу тональності, емоцій та векторизації тексту, що дозволяє глибше аналізувати та ефективніше виявляти фейкові новини, опираючись на унікальні лінгвістичні та культурні особливості української мови.

6.2 Метод ансамблю класифікаторів для онлайн виявлення дезінформації

Нижче представлено розроблений комплексний метод для виявлення дезінформації, який охоплює весь процес – від збору та попередньої обробки даних до аналізу тональності, емоцій та векторизації тексту. Метод використовує передові техніки машинного навчання та нейронних мереж, що забезпечує детальний аналіз текстових даних і ефективне виявлення фейкових новин. На рисунку 6.1 наведено структуру запропонованого методу, яка включає такі етапи:

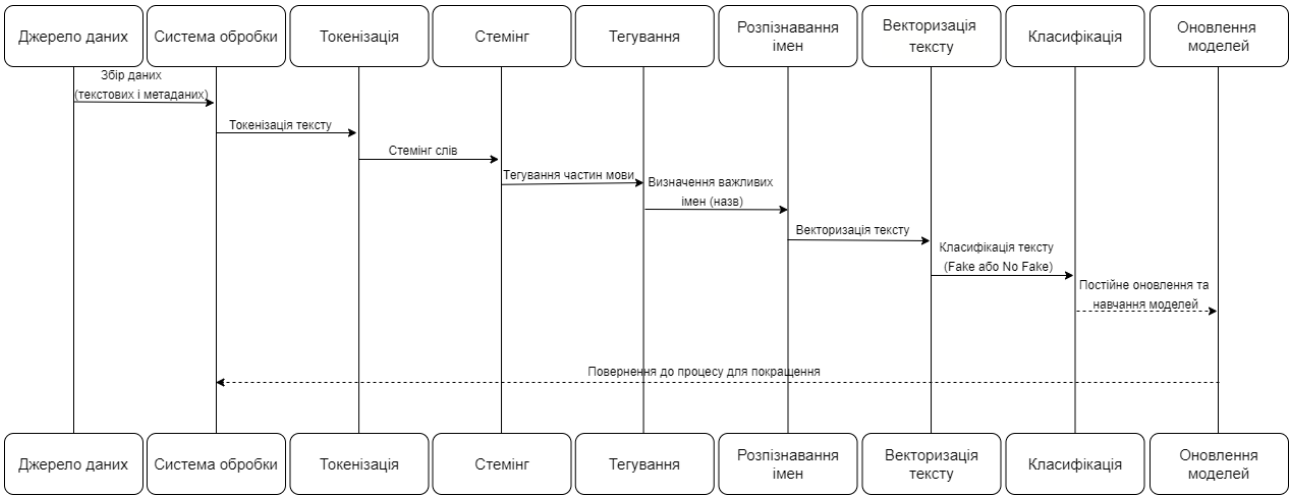


Рисунок 6.1. Структура інтелектуального методу виявлення дезінформації в онлайн режимі

Крок 1. Збір даних. Збір якісних даних є критично важливим етапом у створенні методу для виявлення дезінформації, оскільки він забезпечує ефективне навчання моделей машинного навчання та глибокий аналіз. Детальніше розглянемо цей етап:

1.1. Збір текстових даних.

1.1.1. Визначення джерел. Вибір джерел, таких як новинні портали, соціальні мережі, блоги, форуми та інші платформи, де користувачі обмінюються інформацією.

1.1.2. Автоматизація збору [28]. Використання скриптів або наявних інструментів (веб-скрапінг, API-запити) для автоматизованого збору даних.

1.1.3. Фільтрація та валідація. Відбір і перевірка даних для забезпечення якості та відповідності вимогам дослідження.

1.2. Збір метаданих.

1.2.1. Інформація про авторів. Збір даних про авторів текстів (профілі, історія публікацій, соціальні показники тощо).

1.2.2. Інформація про джерела. Збір додаткових даних (URL, дата публікації, кількість переглядів, лайків, коментарів тощо) для аналізу контексту та популярності.

1.2.3. Структурування метаданих. Організація метаданих у базу даних для зручного доступу та подальшого аналізу.

Крок 2. Попередня обробка даних. Попередня обробка є ключовим етапом у процесі аналізу текстової інформації, що включає різні техніки для підготовки зібраних даних до подальшого аналізу. Ось детальний опис цього кроку [29, 28]:

2.1. Токенізація [30]. Розбиття тексту на окремі слова, фрази, символи або інші значущі елементи, які називаються токенами. Це спрощує подальший аналіз і обробку тексту.

2.2. Стемінг [31]. Видалення суфіксів, префіксів та інфіксів зі слів для зведення їх до основної форми. Це допомагає виявляти загальні теми та закономірності у тексті.

2.3. Тегування частин мови. Визначення граматичних категорій для кожного слова в тексті, що сприяє синтаксичному аналізу та встановленню семантичних зв'язків між словами.

2.4. Розпізнавання іменованих сутностей (Named Entity Recognition). Процес виявлення та класифікації іменованих сутностей у тексті таких, як: імена осіб, організацій, місцезнаходжень та інші значущі категорії.

Крок 3. Векторизація тексту. Перетворення текстових даних у числовий формат, що дозволяє аналізувати та обробляти їх за допомогою методів машинного навчання [31]. Далі розглянемо цей процес детальніше:

3.1. Вибір методу векторизації.

3.1.1. Word2Vec [32]. Модель, що навчає векторні представлення слів у багатовимірному просторі, забезпечуючи близькі векторні представлення для слів, які часто зустрічаються разом.

3.1.2. GloVe. Альтернативний метод векторизації слів, що враховує як локальний, так і глобальний статистичний аналіз текстового корпусу для визначення векторних представлень слів.

3.1.3. BERT. Сучасна модель, яка використовує механізм уваги для визначення взаємозв'язків між словами в тексті, що дозволяє навчати глибокі контекстуальні представлення слів.

3.2. Процес векторизації.

3.2.1. Навчання або завантаження моделей. Моделі можна тренувати на власних даних або скористатися попередньо навченими.

3.2.2. Перетворення тексту. Використання обраної моделі для перетворення кожного слова на векторне представлення.

3.3. Побудова векторних представлень.

3.3.1. Векторизація слів. Отримання векторних представлень для кожного слова в тексті.

3.3.2. Векторизація текстів. Агрегація векторних представлень слів для отримання векторних репрезентацій цілих текстів. Це можна здійснити за допомогою усереднення, сумування або інших методів агрегації.

Крок 4. Навчання за методом онлайн-навчання з ковзним вікном для ансамблів текстових класифікаторів. Цей етап зосереджений на розробці та навчанні моделі класифікації, здатної виявляти фейкову інформацію на основі текстового аналізу та інших ознак. Детальний опис цього процесу виглядає так:

4.1. Створення та навчання ансамблю моделей [27].

4.1.1. Створення ансамблю моделей. Вибір та розробка різних моделей класифікації таких, як: логістична регресія, SVM, LSTM, трансформери тощо.

4.1.2. Навчання моделей. Кожна модель окремо навчається на векторизованих текстових даних.

4.1.3. Формування метамоделі. Використання алгоритму, такого як Adam, для оптимізації ваг моделей у складі ансамблю, що дозволяє покращити інтеграцію та вибір прогнозів від кожної окремої моделі.

4.2. Впровадження методу "ковзного вікна" для онлайн-навчання.

4.2.1. Імплементація "ковзного вікна". Визначення розміру "ковзного вікна" для вибору останніх текстових документів, які будуть використовуватися для постійного оновлення та навчання моделей.

4.2.2. Онлайн-оновлення моделей. Регулярне оновлення моделей у складі ансамблю, використовуючи нові дані та відкидаючи застарілі, що дозволяє підтримувати актуальність та високу точність прогнозів.

4.2.3. Адаптація до змін у даних. Постійне пристосування метамоделі до змін у текстових даних, що забезпечує ефективне реагування на динаміку та нестаціонарність текстових потоків.

Крок 5. Адаптація та перенавчання. Цей етап спрямований на забезпечення здатності системи адаптуватися до змін та еволюції форм дезінформації, що дозволяє підтримувати її ефективність у боротьбі з фейковими новинами [32]. Розгляньмо деталі цього кроку:

5.1. Перенавчання системи.

5.1.1. Збір нових даних. Безперервний збір нових даних з відкритих джерел для врахування актуальних тенденцій та шаблонів дезінформації.

5.1.2. Оцінка потреби в перенавчанні. Аналіз поточної ефективності системи для визначення необхідності перенавчання на основі нових даних.

5.1.3. Перенавчання моделі. Виконання процесу навчання моделі з використанням нових даних, що дозволяє адаптувати її до нових видів дезінформації.

5.2. Оновлення моделей та алгоритмів.

5.2.1. Аналіз нових алгоритмів та технологій. Оцінка та вивчення нових алгоритмів і технологій, які можуть сприяти підвищенню ефективності системи.

5.2.2. Оновлення алгоритмів. Внесення змін до алгоритмів та методів на основі отриманих даних та аналізу результатів.

5.2.3. Тестування та валідація оновлених моделей. Проведення тестування та валідації оновлених моделей для забезпечення їхньої ефективності та надійності.

5.3. Моніторинг та оцінка.

5.3.1. Моніторинг ефективності системи. Постійне спостереження за ефективністю роботи системи для виявлення можливих проблем або напрямків для вдосконалення.

5.3.2. Зворотний зв'язок та адаптація. Збір і аналіз відгуків від користувачів та експертів для подальшого вдосконалення та адаптації системи.

6.3 Реалізація

Для реалізації запропонованого методу було використано набір даних, доступний на Kaggle[34]. Це унікальна колекція, яка налічує близько 60 тисяч заголовків новин, зібраних у період з 24 лютого по 11 грудня 2022 року, що охоплює час повномасштабної російсько-української війни. Набір містить як достовірні, так і фейкові новини, отримані з українських телеграм-каналів та російських каналів з дезінформацією, що робить його найбільшим відкритим джерелом подібних даних. У наборі представлені два основні атрибути: текст заголовку новини та мітка, яка вказує на достовірність (True для підтверджених новин і False для фейкових). Загалом, у наборі міститься 4522 записи з міткою 'False'. Джерела даних включають такі телеграм-канали, як "СУСПІЛЬНЕ НОВИНИ", "Perepichka NEWS" та інші.

На початковому етапі реалізації дані були підготовлені шляхом перетворення текстових матеріалів у числовий формат за допомогою TF-IDF векторизації. На основі цих векторизованих даних були натреновані кілька окремих моделей, зокрема логістична регресія, SVM, випадковий ліс, градієнтний бустинг, KNN, дерево рішень, XGBoost та AdaBoost. Кожну з моделей було налаштовано та оптимізовано для виконання класифікації на навчальному наборі даних. На другому етапі було створено метамодель на основі XGBoost, яку навчали прогнозувати за допомогою методики стекінгу на основі результатів індивідуальних моделей. Оцінка точності класифікації проводилася на тестовому наборі даних, де результати представлено у вигляді звіту про класифікацію та матриці

помилок, відображеної у форматі теплової карти. Такий підхід демонструє значний потенціал ансамблевих методів та стекінгу для покращення продуктивності машинного навчання в складних задачах класифікації тексту, зокрема у виявленні дезінформації.

Результати класифікації представлені у вигляді звіту (Рисунок 6.2), що містить такі метрики, як точність (precision), відновлення (recall), f1-бал (f1-score) та підтримка (support) для двох категорій: "False" (фейкові новини) та "True" (правдиві новини). Під час оцінки ефективності моделі класифікації новин було проведено аналіз її здатності розрізняти правдиві та фейкові повідомлення. Згідно з отриманими результатами, загальна точність моделі складає 93%, що свідчить про її високу ефективність. Для класу "False", що позначає фейкові новини, точність (precision) склала 0.95. Це свідчить про високу ймовірність, що новина справді є фейковою, якщо модель класифікує її як таку. Однак показник відновлення (recall) для цього класу становить 0.72, що означає, що модель не виявила близько 28% фейкових новин. Для класу "True", що відповідає правдивим новинам, модель досягла високої повноти на рівні 0.99, успішно виявляючи 99% правдивих новин. Така здатність точно розпізнавати достовірну інформацію є важливою в боротьбі з дезінформацією.

Аналітичний підхід до оцінки ефективності моделі також включав аналіз матриці помилок. Згідно з результатами, модель правильно ідентифікувала 1793 випадки як істинно-негативні (True Negative), що вказує на успішну класифікацію фейкових новин. Однак було виявлено 705 хибно-негативних результатів (False Negative), що свідчить про те, що модель помилково класифікувала частину фейкових новин як правдиві. Для категорії правдивих новин зафіксовано 8138 істинно-позитивних (True Positive) та 99 хибно-позитивних (False Positive) випадків, що демонструє високу точність у визначенні правдивого контенту. F1-оцінка, яка поєднує

точність і повноту, склала 0.82 для класу "False" та 0.95 для класу "True", що свідчить про загальну збалансованість та надійність моделі.

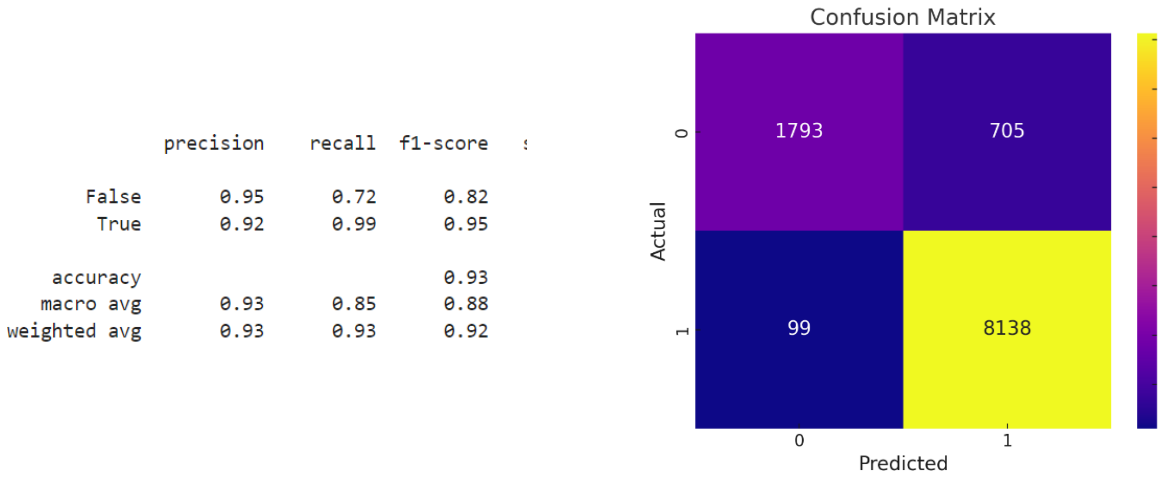


Рисунок 6.2. Результати оцінювання

Аналізуючи результати, наведені в Таблиці 6.1, можна зробити науково обґрунтовані висновки про ефективність методу ECOD у порівнянні з традиційними підходами класифікації. Розглядаючи кожну метрику окремо, видно, що метод ECOD демонструє значні покращення за більшістю показників.

Таблиця 6.1.

Порівняння з розробленим методом

Model	Precision (False)	Precision (True)	Recall (False)	Recall (True)	F1-score (False)	F1-score (True)	Accuracy	Confusion Matrix (False, True)
Logistic Regression	0.96	0.85	0.44	0.99	0.60	0.92	86.42%	(1091, 1407), (51, 8186)
SVM	0.90	0.93	0.75	0.99	0.85	0.96	90.61%	(1680, 818), (68, 8169)
Random Forest	0.88	0.95	0.81	1.00	0.89	0.97	90.31%	(2026, 472), (31, 8206)
Gradient Boosting	0.94	0.86	0.48	0.99	0.64	0.92	87.31%	(1209, 1289), (73, 8164)
KNN	0.91	0.87	0.51	0.99	0.66	0.92	87.57%	(1285, 1213), (121, 8116)
Decision Tree	0.81	0.95	0.83	0.94	0.82	0.94	91.45%	(2077, 421), (497, 7740)
XGBoost	0.89	0.89	0.61	0.98	0.73	0.93	89.19%	(1533, 965), (195, 8042)
AdaBoost	0.88	0.87	0.50	0.98	0.64	0.92	86.92%	(1259, 1239), (165, 8072)
OLTW-TEC	0.95	0.92	0.72	0.99	0.82	0.95	93%	(1793, 705), (99, 8138)

У контексті точності (precision) для класу "False", ECOD досягає значення 0.95, що є порівнянним з результатами логістичної регресії та перевершує більшість інших класичних методів, за винятком Gradient Boosting. Це підкреслює високу здатність ECOD правильно ідентифікувати фейкові новини. Щодо класу "True", точність ECOD становить 0.92, що є конкурентним показником, особливо у порівнянні з методами SVM та XGBoost.

Показник відновлення (recall) для класу "False" у методі ECOD становить 0.72, що значно перевищує аналогічний показник у логістичної регресії та дорівнює результатам SVM, що свідчить про покращену здатність моделі виявляти фейкові новини серед негативного класу. Для класу "True" показник відновлення становить 0.99, що є загальною тенденцією для всіх розглянутих моделей, підкреслюючи їх здатність до виявлення правдивих новин.

Гармонійне середнє між точністю та відновленням, відоме як F1-оцінка (Рисунок 6.3), у методі ECOD дорівнює 0.82 для класу "False" та 0.95 для класу "True" (Рисунок 6.4), що свідчить про збалансованість моделі за цими двома важливими показниками. Це значно перевищує F1-оцінку логістичної регресії для класу "False" та відповідає найкращим результатам серед інших класичних методів.

Загальна точність ECOD становить 93%, що є одним із найвищих показників серед усіх розглянутих моделей, підтверджуючи її ефективність як надійного інструмента для класифікації новин. Матриця помилок також демонструє велику кількість правильно класифікованих прикладів для обох класів, що підкреслює високу точність моделі.

Рисунок 6.3. Оцінки точності

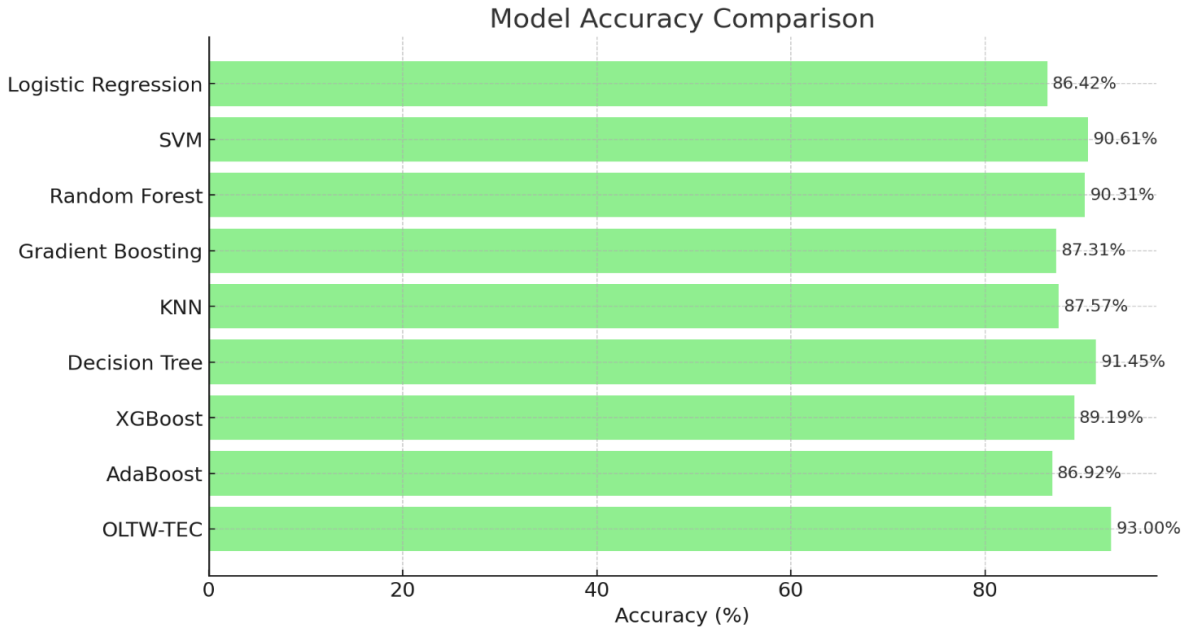


Рисунок 6.4. Оцінка точності прогнозу

6.4 Висновки з проведеного дослідження

Під час проведеного дослідження було ретельно проаналізовано та оцінено ефективність методу ECOD (Ансамбль класифікаторів для онлайн виявлення дезінформації), який є інноваційним підходом до задач класифікації тексту в умовах змінних даних і динамічного онлайн-навчання. Цей метод включає використання ансамблю адаптивних класифікаторів, векторизації текстів, оптимізацію ваг класифікаторів за

допомогою алгоритму Adam, а також впровадження механізму "ковзного вікна" для підтримки актуальності моделі.

Згідно з результатами оцінювання, представленими у звіті класифікації, ECOD продемонстрував високу точність, досягнувши показника 93%. Для класу "False" точність і відновлення склали 0.95 та 0.72 відповідно, що свідчить про ефективність виявлення фейкових новин із мінімальною кількістю пропусків. Для класу "True" модель забезпечила майже ідеальне відновлення на рівні 0.99, підтверджуючи свою здатність до розпізнавання правдивих новин. Отримані результати показують, що ECOD має високу збалансованість між різними метриками, забезпечуючи надійну та ефективну класифікацію в реальних умовах.

Аналіз матриці помилок свідчить про високу точність класифікації, з мінімальною кількістю хибно-позитивних та хибно-негативних результатів. Це підкреслює надійність ECOD як інструмента для фільтрації інформації, що особливо важливо в боротьбі з дезінформацією.

Порівняно з традиційними методами класифікації, ECOD демонструє кращі результати за більшістю метрик і пропонує можливість адаптації до змін у характеристиках даних. Регулювання розміру "ковзного вікна" в залежності від специфіки даних забезпечує додаткову гнучкість та точність методу.

Список використаних джерел:

1. Tao, W., & Peng, Y. (2023). Differentiation and unity: A Cross-platform Comparison Analysis of Online Posts' Semantics of the Russian–Ukrainian War Based on Weibo and Twitter. Communication and the Public, 205704732311655. https://doi.org/10.1177/20570473231165563

2. Mainych, S., Bulhakova, A., & Vysotska, V. (2023). Cluster Analysis of Discussions Change Dynamics on Twitter about War in Ukraine. Proceedings of

the 7th International Conference on Computational Linguistics and Intelligent Systems. Volume II: Computational Linguistics Workshop Kharkiv, Ukraine, Vol-3396, 490–530. https://ceur-ws.org/Vol-3396/paper39.pdf

3. Vasist, P. N., & Krishnan, S. (2023). Fake news and sustainability-focused innovations: A review of the literature and an agenda for future research. Journal of Cleaner Production, 135933. https://doi.org/10.1016/j.jclepro.2023.135933

4. Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon, Article e20382. https://doi.org/10.1016/j.heliyon.2023.e20382

5. Kondamudia, M. R., Sahoob, S. R., Chouhanc, L., & Yadavd, N. (2023). A Comprehensondamudia, M. R., Sahoob, S. R., Chouhanc, L., & Yadavd, N. (2023). A Comprehensive survey of Fake news in Social Networks: Attributes, Features, and Detection Approaches. Journal of King Saud University - Computer and Information Sciences, 101571. https://doi.org/10.1016/j.jksuci.2023.101571

6. Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. AI Open. https://doi.org/10.1016/j.aiopen.2022.09.001

7. Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. Applied Soft Computing, 110235. https://doi.org/10.1016/j.asoc.2023.110235

8. Das, B., & TSB, S. (2023). Multi-contextual learning in disinformation research: A review of challenges, approaches, and opportunities. Online Social Networks and Media, 34-35, 100247. https://doi.org/10.1016/j.osnem.2023.100247

9. Ruffo, G., Semeraro, A., Giachanou, A., & Rosso, P. (2023). Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. Computer Science Review, 47, 100531. https://doi.org/10.1016/j.cosrev.2022.100531

10. Baker, M., Jihad, K., & Taher, Y. (2023). Prediction of People Sentiments on Twitter using Machine Learning Classifiers During Russian Aggression in Ukraine. Jordanian Journal of Computers and Information Technology, 1. https://doi.org/10.5455/jjcit.71-1676205770

11. Chang, Q., Li, X., & Duan, Z. (2024). Graph Global Attention Network with Memory: A Deep Learning Approach for Fake News Detection. Neural Networks, 106115. https://doi.org/10.1016/j.neunet.2024.106115

12. Peng, L., Jian, S., Kan, Z., Qiao, L., & Li, D. (2024). Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. Information Processing & Management, 61(1), 103564. https://doi.org/10.1016/j.ipm.2023.103564

13. Ahammad, T. (2024). Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach. Natural Language Processing Journal, 100053. https://doi.org/10.1016/j.nlp.2024.100053

14. Qu, Z., Meng, Y., Muhammad, G., & Tiwari, P. (2023). QMFND: A quantum multimodal fusion-based fake news detection model for social media. Information Fusion, 102172. https://doi.org/10.1016/j.inffus.2023.102172

15. Farhangian, F., Cruz, R. M. O., & Cavalcanti, G. D. C. (2024). Fake news detection: Taxonomy and comparative study. Information Fusion, 103, 102140. https://doi.org/10.1016/j.inffus.2023.102140

16. Fang, X., Wu, H., Jing, J., Meng, Y., Yu, B., Yu, H., & Zhang, H. (2024). NSEP: Early fake news detection via news semantic environment perception. Information Processing & Management, 61(2), 103594. https://doi.org/10.1016/j.ipm.2023.103594

17. Soga, K., Yoshida, S., & Muneyasu, M. (2023). Exploiting stance similarity and graph neural networks for fake news detection. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2023.11.019

18. Yang, H., Zhang, J., Zhang, L., Cheng, X., & Hu, Z. (2024). MRAN: Multimodal relationship-aware attention network for fake news detection. Computer Standards & Interfaces, 89, 103822. https://doi.org/10.1016/j.csi.2023.103822

19. Raja, E., Soni, B., Lalrempuii, C., & Borgohain, S. K. (2024). An adaptive cyclical learning rate based hybrid model for Dravidian fake news detection. Expert Systems with Applications, 241, 122768. https://doi.org/10.1016/j.eswa.2023.122768

20. Luvembe, A. M., Li, W., Li, S., Liu, F., & Wu, X. (2024). CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. Information Processing & Management, 61(3), 103653. https://doi.org/10.1016/j.ipm.2024.103653

21. Jiang, Y., Yu, X., Wang, Y., Xu, X., Song, X., & Maynard, D. (2023). Similarity-Aware Multimodal Prompt Learning for Fake News Detection. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4347542

22. Syed, L., Alsaeedi, A., Alhuri, L. A., & Aljohani, H. R. (2023). Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda. Array, 100309. https://doi.org/10.1016/j.array.2023.100309

23. Xie, B., & Li, Q. (2023). Detecting fake news by RNN-based gatekeeping behavior model on social networks. Expert Systems with Applications, 120716. https://doi.org/10.1016/j.eswa.2023.120716

24. Přibáň, P., Hercig, T., & Steinberger, J. (2019). Machine Learning Approach to Fact-Checking in West Slavic Languages. In Recent Advances in

Natural Language Processing. Incoma Ltd., Shoumen, Bulgaria. https://doi.org/10.26615/978-954-452-056-4_113

25. Bucos, M., & Drăgulescu, B. (2023). Enhancing Fake News Detection in Romanian Using Transformer-Based Back Translation Augmentation. Applied Sciences, 13(24), 13207. https://doi.org/10.3390/app132413207

26. Afanasieva, I., Golian, N., Golian, V., Khovrat, A., & Onyshchenko, K. (2023). Application of Neural Networks to Identify of Fake News. Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems. Volume II: Computational Linguistics Workshop Kharkiv, Ukraine, (Vol-3396), 346–358. https://ceur-ws.org/Vol-3396/paper28.pdf

27. Bodyanskiy, Y. V., Lipianina-Honcharenko, K. V., & Sachenko, A. O. (2022). Ensemble of adaptive predictors for multivariate nonstationary sequences and its online learning. Radio Electronics, Computer Science, Control, (4(67)), 91–97. doi:10.15588/1607-3274-2023-4-9

28. Gramyak, Roman, Hrystyna Lipyanina-Goncharenko, Anatoliy Sachenko, Taras Lendyuk, and Diana Zahorodnia. "Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring." In IntelITSIS, pp. 246-257. 2021. https://ceur-ws.org/Vol-2853/paper31.pdf

29. Concept of the Intelligent Guide with AR Support / K. Lipianina-Honcharenko et al. International Journal of Computing. 2022. P. 271–277. URL: https://doi.org/10.47839/ijc.21.2.2596

30. Intelligent Information System for Product Promotion in Internet Market / K. Lipianina-Honcharenko et al. Applied Sciences. 2023. Vol. 13, no. 17. P. 9585. URL: https://doi.org/10.3390/app13179585

31. Intelligent Method of Forming the HR Management Short-Term Project / H. Lipyanina et al. Advances in Intelligent Systems and Computing. Cham, 2020. P. 1045–1055. URL: https://doi.org/10.1007/978-3-030-63270-0_71

32. Neural Network Approach for Semantic Coding of Words / Golovko, V., Kroshchanka, A., Komar, M., Sachenko, A. Advances in Intelligent Systems and Computing, 2020, 1020, pp. 647–658. URL: https://link.springer.com/chapter/10.1007/978-3-030-26474-1_45.

33. An Intelligent Method for Forming the Advertising Content of Higher Education Institutions Based on Semantic Analysis / K. Lipianina-Honcharenko et al. Communications in Computer and Information Science. Cham, 2022. P. 169–182. URL: https://doi.org/10.1007/978-3-031-14841-5_11

34. Ukrainian news. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news?resource=download

7

ГІБРИДНИЙ ПІДХІД ДО ВИЗНАЧЕННЯ ДІПФЕЙКІВ НА ОСНОВІ ЛЮДСЬКОГО ОБЛИЧЧЯ

Піцун Олег[1], Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Телька Микола [1], Мельник Назар[1],Лук'янчук Василь[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

7.1 Сучасні підходи до визначення діпфейків

З розвитком штучного генеративного інтелекту зростає потреба у розпізнаванні корисного контенту від фейкового. Сьогодні фейкові дані можуть бути представлені не тільки у текстовій формі, але й у вигляді фото чи відео. Наприклад, діпфейки часто використовуються для шахрайства — видавання себе за іншу особу або поширення неправдивої інформації з метою маніпуляції чи пропаганди. Такі технології також можуть слугувати для створення фальшивих відео або зображень для компрометації людей.

Широкий доступ до інструментів, як-от DeepFake та Face2Face, дозволяє будь-кому генерувати фейкові обличчя. Для створення таких діпфейків застосовуються технології глибокого навчання, особливо згорткові нейронні мережі (CNN), які відзначаються високою ефективністю при роботі з відео та зображеннями. Серед популярних архітектур CNN варто виділити AlexNet, LeNet, MobileNet, ResNet та VGG19.

Створення фейкових облич зазвичай передбачає різні способи маніпуляції, як-от зміна розміру зображення, застосування фільтрів, корекція контрасту та яскравості. У більшості випадків використовують

методи маніпуляції, пов'язані зі зміною виразу обличчя чи його характеристик.

Оскільки генеративний інтелект може бути використаний не лише з позитивними, але й з недобрими намірами, важливою задачею стає визначення реальності зображень чи відео. Глибинне навчання часто використовується для створення фейкових медіа, і для їх аналізу розглядаються різні методи, зокрема виявлення слідів обробки зображення.

Однією з популярних методів є просторово насичена модель (SRM) [1], спочатку розроблена для виділення стеганалітичних ознак. Архітектура згорткової нейронної мережі MISLNet, наприклад, застосовує обмежену згортку та адаптивне навчання для відстеження маніпуляцій за допомогою зворотного поширення.

Серед популярних інструментів для маніпуляції обличчям — DeepFakes, Face2Face та FaceSwap, які застосовують технології комп'ютерного зору для точного виділення та обробки елементів обличчя.

Визначено два основні підходи до створення діпфейків:

- маніпуляція виразом обличчя;
- маніпуляція ідентифікацією обличчя.

Виявлення та запобігання діпфейків є ключовим для забезпечення безпеки та захисту прав людини. Наприклад, у роботі [2] запропоновано модель CNN для аналізу виразу обличчя та фреймворк для виділення ознак фейкових зображень. У роботі [3] представлено підхід FakeCatcher, що використовує біологічні сигнали для визначення фальшивок, а [4] пропонує порівняльний аналіз існуючих методів та модель MesoNet на базі CNN. Аналіз існуючих технологій створення та виявлення діпфейків, наведений у роботі [5], дає комплексний огляд методів та наборів даних у цій сфері.

У процесі аналізу діпфейків важливим є пошук методів обробки великих даних, як розглянуто у роботі [6]. Також для обробки облич

застосовують сучасні підходи до сегментації зображень, що включають виділення ключових областей, таких як очі чи губи, як описано у [7].

7.2 Фактори, що визначають діпфейк

Для виявлення діпфейків слід звертати увагу на такі ключові фактори:

- Натуральність шкіри: на діпфейках часто зображена занадто гладка шкіра, тоді як деякі елементи обличчя (наприклад, очі) можуть виглядати старішими або мати невідповідні деталі.
- Фізика тіней: діпфейк-алгоритми можуть не відображати правильне положення та природність тіней. У таких випадках особливо варто звернути увагу на область очей та брів.
- Освітлення: діпфейки можуть некоректно відтворювати фізику освітлення. Варто спостерігати за відблисками на окулярах, перевіряючи, чи змінюється кут відблиску під час руху людини.
- Реалістичність волосся: штучно згенероване волосся на обличчі (борода, вуса) часто виглядає неприродньо або ненатурально однорідним.
- Наявність родимок: перевірка родимок та інших індивідуальних особливостей може вказати на аномалії в зображенні.
- Рух губ: на діпфейках губи іноді рухаються надто синхронно, що робить міміку неприродною та надто «відточеною» [8].

Для створення діпфейків сьогодні найчастіше використовують генеративний штучний інтелект, зокрема генеративні змагальні мережі (GAN). Для визначення діпфейків широко застосовуються методи комп'ютерного зору та згорткові нейронні мережі, а також іноді інші типи нейронних мереж.

Одним із підходів є аналіз статистичних характеристик зображення за допомогою алгоритмів комп'ютерного зору, як це детально розглянуто в роботі [9].

Інший підхід включає розробку нових архітектур згорткових нейронних мереж для якісної класифікації реальних та фейкових зображень. Один з важливих методів виявлення фейків — аналіз частоти кліпання очима у відео, оскільки штучні моделі часто не відтворюють цей природний рух з потрібною точністю.

Крім того, на діпфейках часто спостерігається розмитий простір навколо голови.

7.3 Узагальнений підхід до визначення діпфейків

До основних датасетів для визначення фейків можна віднести такі:

Flickr-Faces-HQ (FFHQ) – вміщує 70,000 зображень в хорошій якості та розмір 1024 на 1024 пікселі. Особливістю даної мережі є наявність фото з різними етнічними та віковими характеристиками, що дозволяє зробити нейронну мережі більш універсальною.

100K-Faces - вміщує 100,000 зображень. Особливість полягає у тому, що зображення згенеровані з допомогою нейронної мережі StyleGAN.

mEBAL – датасет вміщує матеріали, для можливості визначення кількості кліпань очима. Зазвичай цей датасет використовують у задачах оцінки рівня уваги, аналізу нейродегенеративних захворювань, розпізнавання обману та виявлення втоми водія.

VGGFace2 одна із найпотужніших баз, що вміщує більше 3-ох мільйонів зображень. Її особливість полягає у тому, що вона вміщує зображення відомих людей, політиків, акторів тощо.

На рисунку 7.1 наведено узагальнений підхід до визначення діпфейків, що базується на алгоритмах комп'ютерного зору та архітектурах згорткових нейронних мереж.

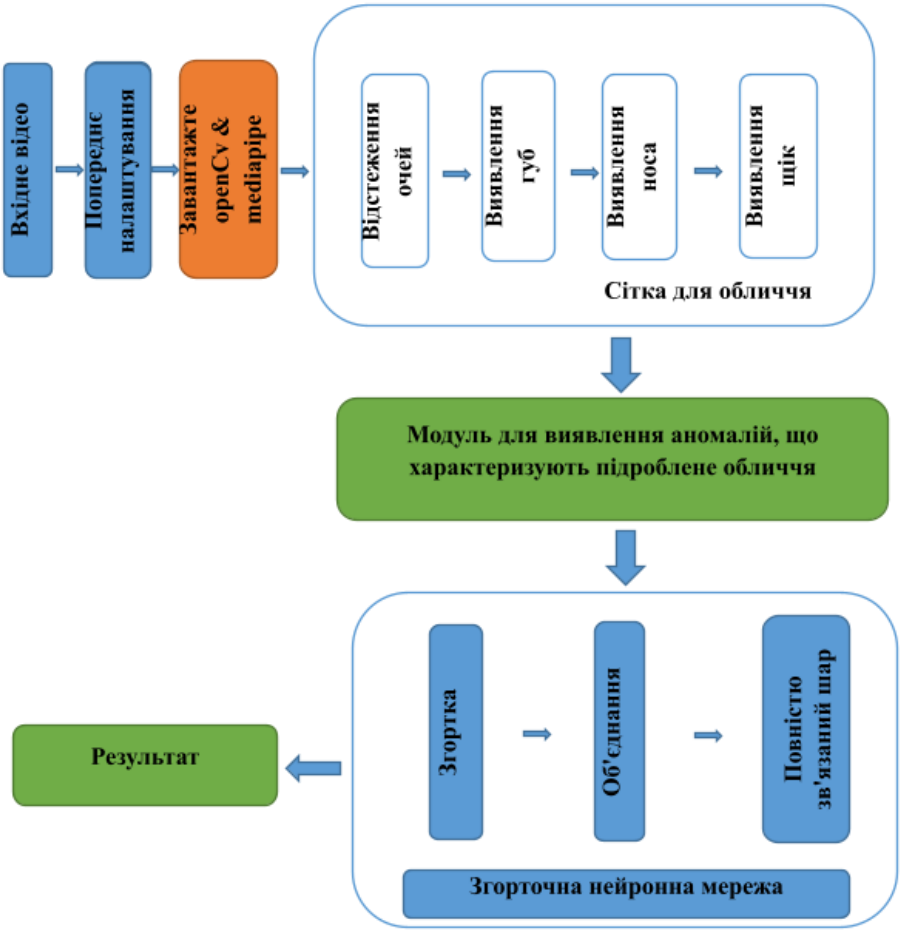


Рисунок 7.1. Узагальнений підхід до визначення діпфейку

На першому етапі відбувається захоплення відеофрагмента з людським обличчям. На наступному етапі здійснюється попередня підготовка відео. Для виділення необхідних елементів на обличчі використовуються сучасні бібліотеки комп'ютерного зору, такі як OpenCV та Mediapipe. Також необхідно налаштувати середовище розробки Python та підключити бібліотеки для візуалізації та обчислень.

Принцип роботи полягає в тому, що задаються ключові точки на обличчі людини, які необхідні для ідентифікації різних частин обличчя, таких як очі, брови, губи, щоки тощо. Ці точки подаються у вигляді масиву

координат. Далі, в режимі реального часу, здійснюється відслідковування положення точок та обчислення відстаней між ними. Наприклад, для визначення кількості кліпань очей використовують евклідову відстань між точками на повіках.

Крім того, за допомогою точок на обличчі можна вимірювати відстань між краями губ, що дозволяє визначити, чи людина посміхається. Це також дозволяє оцінити інші аспекти, такі як симетрія обличчя. Детекція щік або лоба важлива для подальшого аналізу кольору шкіри. Якщо тон шкіри надто ідеальний або рівномірний, це може свідчити про наявність діпфейків.

На фінальному етапі застосовуються згорткові нейронні мережі та інші архітектури нейронних мереж для визначення приналежності зображення до класу реальних або фейкових об'єктів.

Завдяки такому комбінованому підходу формується кінцеве рішення про реальність обличчя на відео чи зображенні.

Використання бібліотеки Mediapipe для задачі виділення ключових точок на обличчі наведено на рисунку 7.2.

Рисунок 7.2. Виділення ключових точок на обличчі

В результаті використання такого підходу можна визначити необхідні точки, наприклад, навколо очей чи губ для можливості подальшого обчислення їхніх характеристик.

Важливою складовою будь-якої системи, що використовує штучний інтелект та великі набори даних є можливість їхнього використання в хмарних середовищах. В таких випадках важливим аспектом є використання MLOps – підходів. Приклад конфігураційного файлу наведено на рисунку 7.3.

```
resource "digitalocean_droplet" "fake_detection" {
  ssh_keys = [
    digitalocean_ssh_key.default.fingerprint
  ]
  image     = "ubuntu-20-04-x64"
  name      = " fake_detection "
  region    = "nyc1"
  size      = "s-1vcpu-1gb"
  user_data = file("fake_detection_app.yaml")

  connection {
    host = self.ipv4_address
    user = "root"
    type = "ssh"
    private_key = file("tf-digitalocean")
    timeout = "2m"
  }
  provisioner "remote-exec" {
    inline = [
      "export PATH=$PATH:/usr/bin",
      # install nginx
      "sudo apt-get update",
      "sudo mkdir experiments",
      "sudo cd experiments",
      "sudo unzip link_to_dataset",
      "sudo git clone link_to_project"
    ]
  }
```

Рисунок 7.3. Приклад конфігураційного файлу terraform

Для можливості реалізації розгортання проекту на хмарному середовищі вибрано технологію terraform, що дозволяє використовувати підхід IaC для спрощення процесу інтегрування коду .

Варто відзначити, що в такому випадку значну увагу приділяють процесу завантаження датасету та самого проекту на хмарне сховище.

7.4 Висновки з проведеного дослідження

На основі аналітичного підходу проведено аналіз існуючих рішень, інструментів, датасетів для визначення сучасних тенденцій у розробці систем для визначення діпфейків на основі аналізу людського обличчя.

У роботі запропоновано узагальнений підхід до класифікації діпфейків з використанням як алгоритмів комп'ютерного зору так і з використанням згорткових нейронних мереж, навчених на основі відомих датасетів.

Перевагою запропонованого підходу є те, що для оцінки реальності чи фейковості зображень використовуються не лише нейронні мережі, які вимагають значних обчислювальних ресурсів, таких як GPU і хмарні технології, а й алгоритми комп'ютерного зору. Це дозволяє точково досліджувати окремі елементи обличчя для подальшого аналізу та виявлення характеристик, що можуть свідчити про наявність діпфейків.

Крім того, в роботі запропоновано структуру terraform-файлу для розгортання проекту в хмарному середовищі, що забезпечує зберігання датасету та необхідні ресурси для обробки.

Список використаних джерел:

1. Goljan M. and Fridrich J., ‘‘CFA-aware features for steganalysis of color images,’’ Proc. SPIE, vol. 9409, Feb. 2015, Art. no. 94090V
2. Kim, E., & Cho, S. (2021). Exposing fake faces through deep neural networks combining content and trace feature extractors. IEEE Access, 9, 123493-123503.
3. Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE transactions on pattern analysis and machine intelligence.
4. https://arxiv.org/pdf/1809.00888v1

5. Almars, A. (2021) Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications, 9, 20-35. doi: 10.4236/jcc.2021.95003.

6. Xuan, X., Peng, B., Wang, W. and Dong, J. (2019) On the Generalization of GAN Image Forensics. In: Chinese Conference on Biometric Recognition, Springer, Berlin, 134-141. https://doi.org/10.1007/978-3-030-31456-9_15

7. ISDMCI 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 149. Springer, Cham. https://doi.org/10.1007/978-3-031-16203-9_28

8. Berezsky O., Melnyk G., Batko Y. and Pitsun O., "Regions matching algorithms analysis to quantify the image segmentation results," 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2016, pp. 33-36, doi: 10.1109/STC-CSIT.2016.7589862

9. https://www.media.mit.edu/projects/detect-fakes/overview/

10.Berezsky, O., Pitsun, O., Liashchynskyi, P., Derysh, B., Batryn, N. (2023). Computational Intelligence in Medicine. In: Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making.

8

ПОРІВНЯЛЬНИЙ АНАЛІЗ АРХІТЕКТУР CNN ДЛЯ КЛАСИФІКАЦІЇ ЕМОЦІЙНИХ СТАНІВ НА ЛЮДСЬКОМУ ОБЛИЧЧІ

Піцун Олег[1], Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Телька Микола[1], Юрків Христина[1], Лендюк Дмитро [1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

8.1 Порівняння архітектур ЗНМ для класифікації емоцій за зображеннями обличчя

Емоції є невід'ємною частиною нашого життя, допомагаючи покращити взаємодію між людьми та зрозуміти ставлення конкретної особи до певної ситуації. Вміння розпізнавати емоції дозволяє глибше зрозуміти співрозмовника або групу співрозмовників. Штучний інтелект, який сьогодні відіграє важливу роль у всіх сферах життя, також здатен покращувати якість життя, тому його застосування для оцінки емоційного стану людини є актуальним завданням.

Для розпізнавання зображень, зокрема обличь, зараз використовуються згорткові нейронні мережі, які зарекомендували себе як найкращий підхід для класифікації графічних даних порівняно з іншими методами. Існує багато архітектур згорткових мереж, але вони демонструють високу точність класифікації переважно для специфічних типів зображень, і універсального підходу до класифікації всіх зображень досі не існує. Використання таких нейронних мереж також допомагає

усунути суб'єктивні фактори в оцінці емоційного стану людини. Крім того, автоматичне розпізнавання емоцій дозволяє виявляти фейкові стани, що є актуальним у зв'язку з розвитком технологій створення штучного контенту.

Важливим аспектом створення якісної системи класифікації є добір вибірки для навчання та тестування. У цьому дослідженні використовується датасет CK+48, який містить достатню кількість класів та зображень для проведення аналізу. До того ж, цей датасет широко застосовується в наукових дослідженнях, що дозволяє відтворити проведені експерименти з високою точністю.

8.2 Методи класифікації емоцій за зображеннями облич

Науковці значну увагу приділяють підходам до класифікації та розпізнавання емоцій на обличчях. Для цього розроблено декілька популярних датасетів, серед яких — CK+48 [1]. У роботі [2] представлено аналіз емоцій на обличчях на основі датасету зі зображеннями студентів. Автори пропонують використовувати додаткові сенсори для підвищення точності визначення емоційного стану.

У дослідженні [3] розглянуто різні підходи до визначення емоцій на обличчі на основі трьох різних датасетів. Опис, структура та процес формування датасету CK+48 наведено у роботі [4]. Робота [5] присвячена розробці підходів до визначення емоцій на обличчях у режимі реального часу для шести основних емоційних станів. Узагальнений огляд підходів до розпізнавання людських емоцій і класифікації з використанням нейронних мереж представлено в роботах [6,7].

У дослідженні [8] автори провели розпізнавання емоцій із використанням датасету FER2013, застосовуючи згорткові мережі, зокрема архітектуру VGG, і досягли точності близько 70%. У роботі [9] наведено

архітектуру згорткової нейронної мережі для розпізнавання емоцій у реальному часі.

Аналіз застосування елементів штучного інтелекту та систем комп'ютерного зору для обробки зображень представлено в роботах [10-12]. Зокрема, розглянуто підходи, що дозволяють прискорити обробку зображень, та запропоновано нові архітектури згорткових нейронних мереж для класифікації специфічних типів зображень.

8.3 Узагальнений алгоритм класифікації зображень

В якості датасету для даної роботи обрано CK+48, що включає до 1000 зображень, поділених на сім класів: angry, happy, fear, neutral, disgust, sad та surprise [13]. Датасет складається із зображень розміром 48 x 48 пікселів. Для розподілу на тестову та навчальну вибірки було здійснено поділ зображень в окремі директорії «test» та «training». Узагальнений алгоритм класифікації зображень емоцій на людському обличчі представлений на рисунку 8.1.

Рисунок 8.1. Узагальнений алгоритм класифікації зображень емоцій на людському обличчі

Узагальнений алгоритм класифікації емоцій включає дві основні складові: формування датасету та навчання нейронної мережі. Якість класифікації та ефективність навчання залежать від правильного розподілу

даних у датасеті та ретельного вибору архітектури згорткової нейронної мережі.

8.4 Аналіз архітектур згорткових нейронних мереж

Згорткові нейронні мережі активно використовуються для класифікації різного роду зображень. Їхня перевага полягає у можливості використовувати зображення як вхідні дані без необхідності додаткового перетворення. Основні компоненти згорткових нейронних мереж:

- Згорткові шари — головний елемент нейронної мережі, що здійснює згортку вхідного зображення з маскою, утворюючи менш деталізовані зображення і зберігаючи ключові ознаки.
- Шар пулінгу — дозволяє зменшити просторовий розмір згорнутої функції, що допомагає виділяти ознаки зображення з більшою точністю.
- Повністю підключені шари — використовуються на фінальному етапі і представляють собою стандартну структуру нейронної мережі для прийняття остаточних рішень.

Важливим параметром при роботі з нейронною мережею є кількість фільтрів, яка впливає на якість роботи мережі, зокрема на здатність виділяти ознаки. Однак надмірна кількість фільтрів може спричинити перенавчання і погіршити результати.

Розмір фільтра визначає, наскільки деталізованою буде функція виділення ознак: більші фільтри підходять для грубих ознак, тоді як менші – для тонших деталей.

Крок (strides) — більший крок зменшує просторовий розмір виходу на наступному шарі, що допомагає контролювати обсяг обчислень.

Функція активації — найбільш популярні функції активації, такі як ReLU, tanh і sigmoid, дозволяють мережі моделювати нелінійні залежності, що підвищує здатність до розпізнавання складних патернів.

LeNet — ця архітектура згорткової нейронної мережі була однією з перших і добре зарекомендувала себе для класифікації невеликих за розміром зображень. LeNet стала основою для подальших досліджень у сфері згорткових мереж і досі використовується для навчання зображень невеликого розміру.

Узагальнений вигляд архітектури згорткової нейронної мережі наведено на рисунку 8.2.

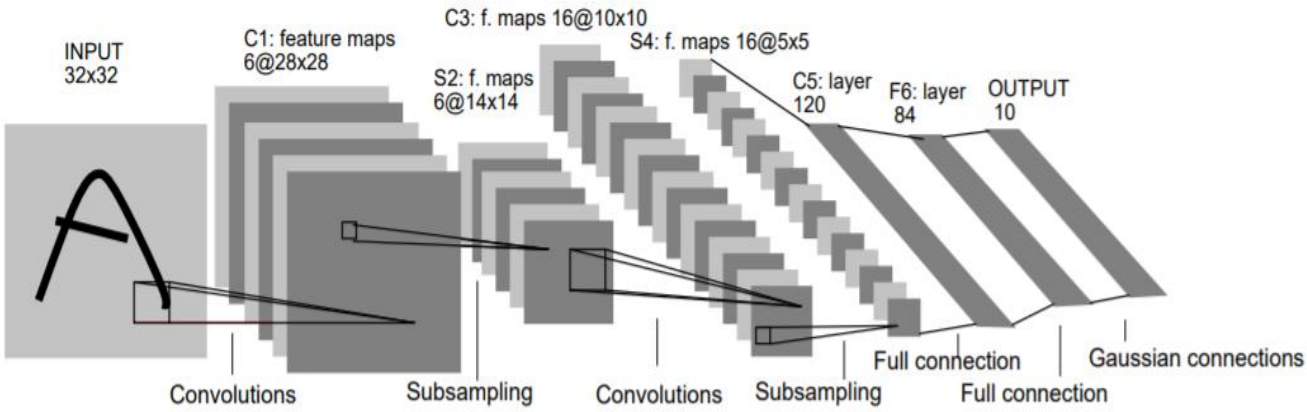


Рисунок 8.2. Узагальнена структура нейронної мережі.

Дану архітектуру активно використовують для розпізнавання дорожніх знаків чи класифікації людських обличь.

AlexNet — одна з найвідоміших архітектур глибоких згорткових нейронних мереж, запропонована у 2012 році. У порівнянні з LeNet, AlexNet має значно глибшу архітектуру з більшою кількістю шарів, що дозволяє виділяти детальніші та складніші об’єкти на зображеннях. Як функцію активації використовують ReLU, що сприяє швидшій та стабільнішій роботі мережі.

MobileNets — архітектура, створена для ефективної роботи на мобільних пристроях, яка забезпечує низьку затримку (low latency) та високу продуктивність. Завдяки легшій структурі та оптимізованим шарам, MobileNets підходить для систем реального часу та часто перевершує такі моделі, як VGG16, у мобільних застосунках та застосуваннях з обмеженими ресурсами.

Графічне представлення блоків згорткової мережі LeNet наведено на рисунку 8.3. Дана архітектура складається із шарів згортки, шарів max-пулінгу, та 3 шарів Dense.

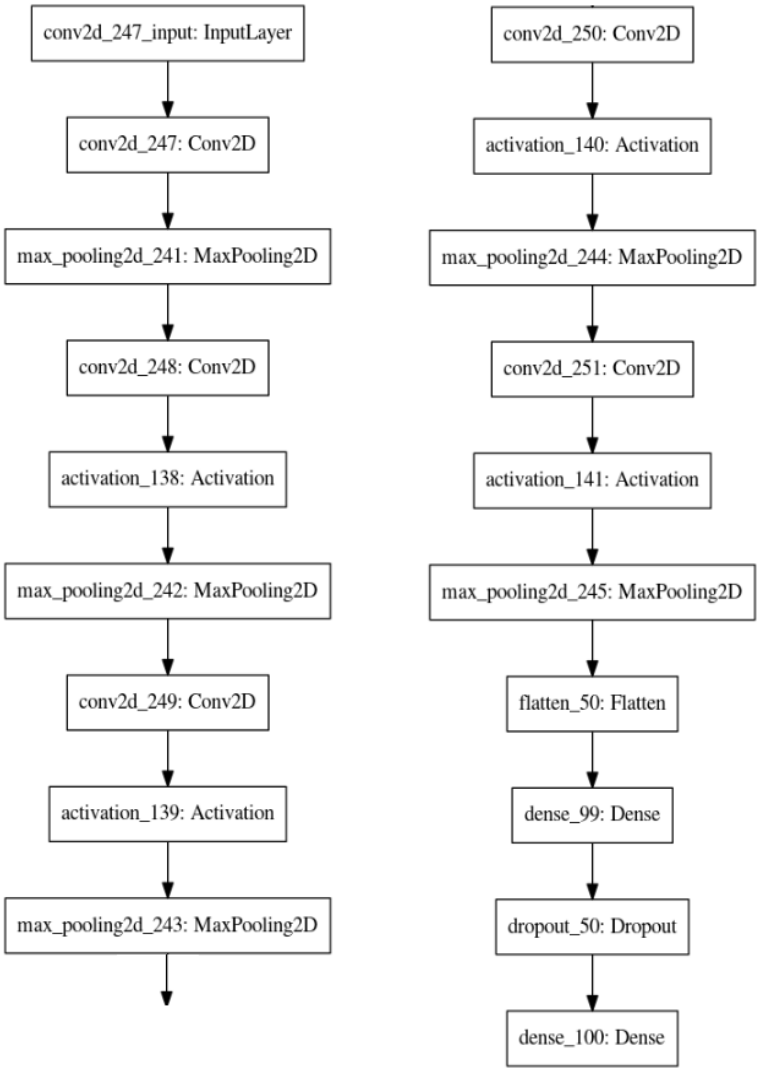


Рисунок 8.3. Графічне представлення запропонованої архітектури

8.5 Комп’ютерні експерименти

Результати використання архітектури AlexNet наведено на рисунку 8.4

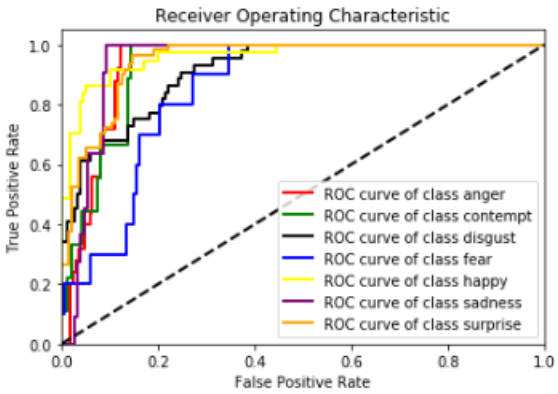


Рисунок 8.4. Результати використання архітектури AlexNet

На даному графіку наведено показники ROC – кривих для 7 класів. Враховуючі наведені дані, можна зробити висновок, що точність класифікації є високою.

Матрицю результатів класифікації наведено на рисунку 8.5.

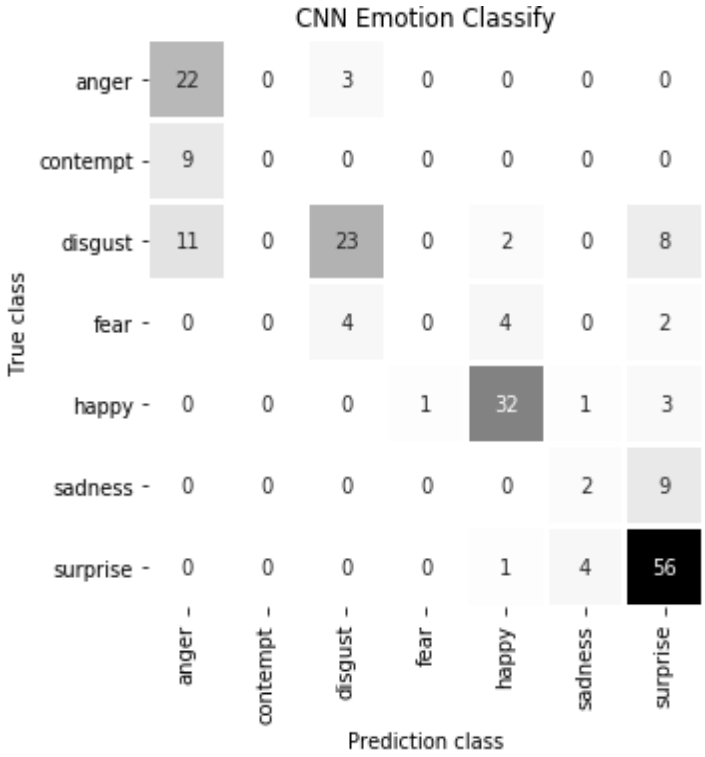


Рисунок 8.5. Матриця плутанини

Параметри шарів для мережі LeNet для зображень з датасету CK+48.

Layer (type)	*Output Shape*	*Param #*
conv2d_73 (Conv2D)	*(None, 48, 48, 16)*	*448*
max_pooling2d_73 (MaxPooling)	*(None, 24, 24, 16)*	*0*
conv2d_74 (Conv2D)	*(None, 24, 24, 16)*	*2320*
activation_37 (Activation)	*(None, 24, 24, 16)*	*0*
max_pooling2d_74 (MaxPooling)	*(None, 12, 12, 16)*	*0*
conv2d_75 (Conv2D)	*(None, 12, 12, 32)*	*12832*
activation_38 (Activation)	*(None, 12, 12, 32)*	*0*
max_pooling2d_75 (MaxPooling)	*(None, 6, 6, 32)*	*0*
conv2d_76 (Conv2D)	*(None, 4, 4, 64)*	*18496*
max_pooling2d_76 (MaxPooling)	*(None, 2, 2, 64)*	*0*
conv2d_77 (Conv2D)	*(None, 2, 2, 64)*	*102464*
activation_39 (Activation)	*(None, 2, 2, 64)*	*0*
max_pooling2d_77 (MaxPooling)	*(None, 1, 1, 64)*	*0*
flatten_19 (Flatten)	*(None, 64)*	*0*
dense_37 (Dense)	*(None, 128)*	*8320*
dropout_19 (Dropout)	*(None, 128)*	*0*
dense_38 (Dense)	*(None, 7)*	*903*

Total params: 145,783

Trainable params: 145,783

Non-trainable params: 0

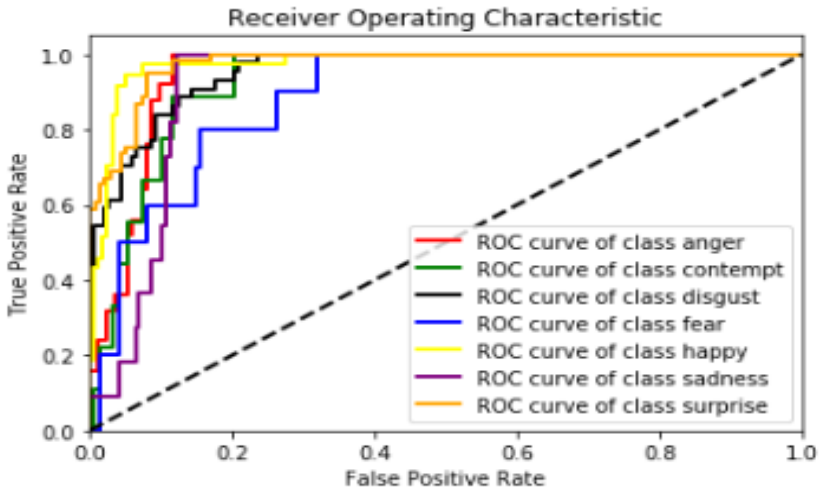


Рисунок 8.6. ROC - крива для спроектованої архітектури

Результати навчання наведено на рисунку 8.7.

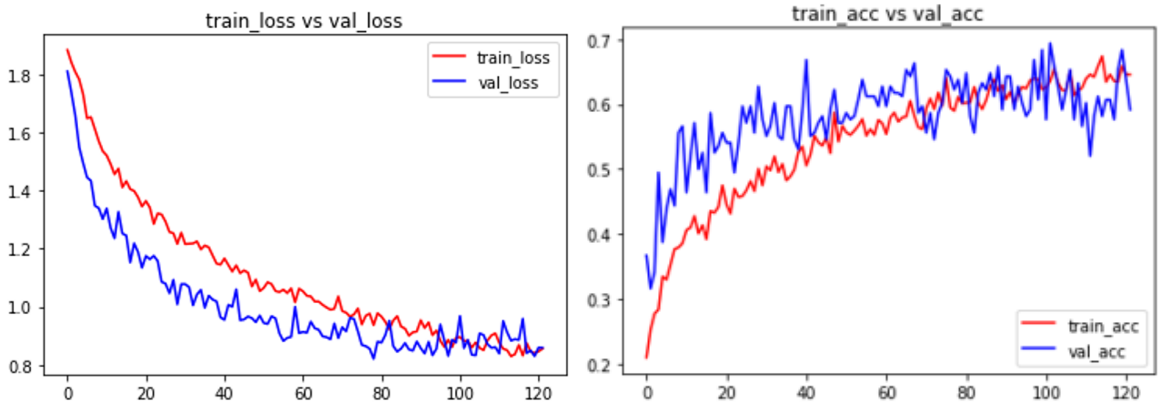


Рисунок 8.7. Результати класифікації нейронних мереж з допомогою архітектури LeNet за параметром Accuracy

ROC – крива для LeNet наведено на рисунку 8.8.

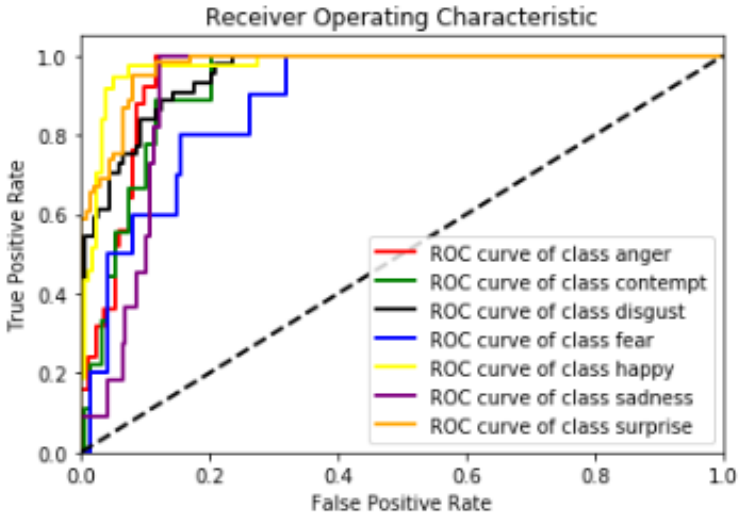


Рисунок 8.8. ROC – крива для розробленої архітектури

Аналізуючи вищенаведені результати можна зробити висновок, що мережа LeNet показала кращі результати у порівнянні з ALexNet

Результати класифікації з допомогою 3 архітектур нейронних мереж наведено на рисунку 8.9.

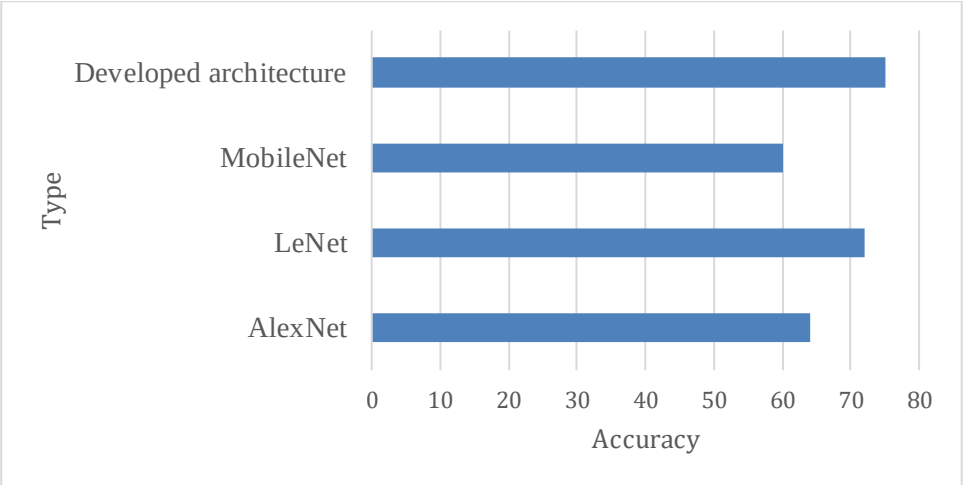


Рисунок 8.9. Результати класифікації

В результаті проведених досліджень на основі датасету CK+48 можна зробити висновок, що архітектури згорткових мереж демонструють приблизно схожі результати, однак, архітектура LeNet продемонструвала кращі результати за критерієм точність – 72% , на відміну від ALexNet чи MobileNet – 64 та 60 % відповідно.

У подальших дослідженнях необхідно розширити мережу архітектур для досліджень та необмежуватись лише найвідомішими.

8.6 Висновки з проведеного дослідження

У межах проведеного дослідження було здійснено порівняльний аналіз ефективності трьох популярних архітектур згорткових нейронних мереж (LeNet, AlexNet та MobileNet) для задачі класифікації емоційних станів на зображеннях людських облич, використовуючи датасет CK+48. Отримані результати засвідчують актуальність використання CNN у розв’язанні задач автоматичного розпізнавання емоцій, що має практичну

цінність для систем безпеки, медицини, освіти та людино-машинної взаємодії.

За результатами комп'ютерних експериментів встановлено, що архітектура LeNet продемонструвала найвищу точність класифікації — 72%, перевершивши AlexNet (64%) та MobileNet (60%). Це підтверджує доцільність застосування саме простіших, але добре налаштованих моделей для класифікації обмежених за розміром і структурою зображень, характерних для CK+48. ROC-криві та матриці плутанини показали стабільну продуктивність LeNet при розпізнаванні більшості з семи базових емоцій.

Розроблений та протестований узагальнений алгоритм обробки зображень, включаючи формування вибірок, попередню обробку та параметризацію моделі, забезпечив високу відтворюваність результатів і може бути адаптований для подальших експериментів з іншими датасетами.

Незважаючи на досягнуту точність, результати дослідження вказують на необхідність подальшого удосконалення архітектур та методик навчання, зокрема розширення спектру тестованих моделей, використання механізмів регуляризації, оптимізації гіперпараметрів, а також включення комбінованих моделей (наприклад, CNN-LSTM) для підвищення загальної продуктивності.

У майбутніх дослідженнях доцільно також дослідити використання мультимодальних підходів, додаткових сенсорів або відеопотоків для розпізнавання емоцій у динаміці, що сприятиме покращенню якості розпізнавання та адаптивності моделей у реальних сценаріях.

Список використаних джерел:

1. Lucey P.,Cohn J. F., Kanade T.,Saragih J.,Ambadar Z., and Matthews I., “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in 2010 ieee computer society conference on computer vision and pattern recognition-workshops, pp. 94–101, IEEE, 2010

2. Magdin, Martin, Benko Ľubomír, and Koprda Štefan. 2019. "A Case Study of Facial Emotion Classification Using Affdex" Sensors 19, no. 9: 2140. https://doi.org/10.3390/s19092140

3. Minaee, Shervin, Mehdi Minaei, and Amirali Abdolrashidi. 2021. "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network" *Sensors* 21, no. 9: 3046. https://doi.org/10.3390/s21093046

4. Lucey P., Cohn J. F., Kanade T., Saragih J., Ambadar Z. and Matthews I., "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

5. Emre DANDIL, and Özdemir Rıdvan. "Real-time Facial Emotion Classification Using Deep Learning Article Sidebar." International Journal of Data Science and Applications 2, no. 1 (2019): 13-17.

6. Moolchandani M., Dwivedi S., Nigam S. and Gupta K., "A survey on: Facial Emotion Recognition and Classification," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1677-1686, doi: 10.1109/ICCMC51019.2021.9418349.

7. Hu, Tianming, Liyanage C. De Silva, and Kuntal Sengupta. "A hybrid approach of NN and HMM for facial emotion classification." Pattern Recognition Letters 23, no. 11 (2002): 1303-1310.

8. Khaireddin, Yousif, and Zhuofa Chen. "Facial emotion recognition: State of the art performance on FER2013." arXiv preprint arXiv:2105.03588 (2021).

9. Duncan, Dan, Gautam Shine, and Chris English. "Facial emotion recognition in real time." Computer Science (2016): 1-7.
10. Berezsky, O., Pitsun, O., Melnyk, G., Koval, V., Batko, Y. (2023). Multi-threaded Parallelization of Automatic Immunohistochemical Image Segmentation. In: Hu, Z., Wang, Y., He, M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics IV. CSDEIS 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 158. Springer, Cham. https://doi.org/10.1007/978-3-031-24475-9_23
11. Berezsky, O., Liashchynskyi, P., Pitsun, O., Liashchynskyi, P., Berezkyy, M. Comparison of Deep Neural Network Learning Algorithms for Biomedical Image Processing. CEUR Workshop Proceeding, 2022, 3302, pp. 135–145. https://ceur-ws.org/Vol-3302/paper7.pdf
12. Pitsun, O. MLOps Approach for Automatic Segmentation of Biomedical Images / Oleh Berezsky, Oleh Pitsun, Grygoriy Melnyk, Yuriy Batko, Petro Liashchynskyi, Mykola Berezkyi // CEUR Workshop Proceeding, 2023, pp. 241–248. https://ceur-ws.org/Vol-3302/paper7.pdf
13. Chaudhari, Aayushi, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. 2022. "ViTFER: Facial Emotion Recognition with Vision Transformers" Applied System Innovation 5, no. 4: 80. https://doi.org/10.3390/asi5040080

9

АНАЛІЗ ЗАСОБІВ ВИЯВЛЕННЯ DEEPFAKE НА ОСНОВІ АНАЛІЗУ ОБЛИЧЧЯ НА ВІДЕО

Піцун Олег[1], Ліп'яніна-Гончаренко Христина[1], Соя Мар'яна[1], Телька Микола[1], Мельник Назар[1],Теліховський Олесь[1]
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

9.1 Сучасні підходи до визначення діпфейків на відео

Маніпуляція мультимедійним контентом завжди була важливою задачею, однак із розвитком генеративного інтелекту значно зросла кількість якісних неправдивих матеріалів, зокрема у вигляді відео та фото. Такий контент має більший вплив на довірливу публіку, порівняно з текстовими повідомленнями. Поняття Deepfake виникло від поєднання термінів "deep learning" та "fake" і використовує алгоритми генеративного інтелекту для створення високоякісних підробок аудіо та відео.

Технології deepfake постійно розвиваються, ускладнюючи створення ефективних методів для виявлення підробок. Алгоритми стають дедалі складнішими, що робить підроблені відео та зображення все реалістичнішими. Оскільки інструменти для створення deepfake стають доступнішими та зручнішими у використанні, навіть непрофесіонали можуть створювати високоякісні підробки. Deepfake використовує штучний інтелект для зміни обличчя людей на зображеннях та відео, синхронізуючи вирази губ, очей, іншу міміку та рухи тіла.

Визначення діпфейків на відео є складною задачею, що вимагає обробки великих обсягів даних та розробки інтелектуальних алгоритмів. Існують підходи, що дозволяють виявляти елементи відео, характерні для діпфейків. Ця галузь активно розвивається і потребує постійного вдосконалення методів виявлення.

Об'єкт дослідження – процес формування та виявлення діпфейків на відео.

Предмет дослідження – інструменти та технології, що використовуються для створення діпфейків на людських обличчях.

Метою даної роботи є аналіз технологій і засобів, що лежать в основі створення діпфейків, а також програмних засобів для їх виявлення.

Для досягнення мети необхідно виконати такі завдання:

- Проаналізувати сучасний стан технологій виявлення діпфейків через огляд наукових публікацій;
- Визначити переваги та обмеження існуючих підходів до створення діпфейків;
- Проаналізувати можливості існуючих інструментів для визначення діпфейків на відео та розробити нові алгоритми й засоби, засновані на штучному інтелекті.

У даній роботі увага зосереджена на комплексному дослідженні методів виявлення глибоких фейків за допомогою методів глибокого навчання, таких як рекурентні нейронні мережі (RNN), згорткові нейронні мережі (CNN) та технології LSTM. Аналіз систем для визначення фейків на відео та фото представлений у роботах [1,2]. Автори описують технології, що використовуються в цих системах, а також отриману точність результатів. Прикладом таких маніпуляцій є популярний додаток для смартфонів FaceApp [3]. Підхід до визначення deepfake на основі підрахунку кількості кліпань розглянуто у роботі [4]. Автори здійснюють

підрахунок кількості кліпань та аналізують їхню частоту і тривалість. Крім того, маніпуляції з атрибутами обличчя можуть змінювати різні аспекти, зокрема колір волосся, шкіри, статі, віку, а також додавання окулярів [5].

У роботі [6] автори провели аналіз існуючих інструментів для формування діпфейків, зокрема розглянули алгоритми комп'ютерного зору, класифікатори, згорткові нейронні мережі, методи опорних векторів, а також метод k найближчих сусідів. У роботі [7] наведено аналіз якості формування діпфейків з використанням різних датасетів, таких як DFFD та Celeb-DF. Аналіз діпфейків проводився на основі згорткових нейронних мереж і показав хороші результати. Гістограма лінійного бінарного шаблону Fisherface з використанням техніки класифікатора DBN (FF-LBPH DBN) була реалізована як метод виявлення зображень deepfake в роботі [8]. Маніпуляції із зображеннями обличчя були проаналізовані за допомогою моделі FF-LBPH DBN, що також забезпечило високий рівень точності.

У роботі [9] автори провели детальний аналіз еволюції DeepFakes, зосереджуючи увагу на обробці областей обличчя та ефективності виявлення фейків. У цьому аналізі використовуються популярні бази даних, такі як UADFV і FaceForensics++ 1-го покоління, а також новітні бази даних, зокрема Celeb-DF і DFDC 2-го покоління. У роботі [10] досліджено чотири нейронні мережі для виявлення глибоких фейків, використовуючи частину наборів даних Celeb-DF, Celeb-DF-v2 і Deep Fake Detection Challenge (DFDC). Автори порівняли ці моделі за їхніми характеристиками, щоб визначити найбільш точну та ефективну серед них. У роботі [11] запропонований метод, що досягає найвищої точності — 99% при проведенні різних типів експериментів на стандартних наборах даних. Також обговорюється узагальнення та пояснюваність запропонованої моделі. Запропонований метод у роботі [12] використовує механізм уваги для обробки карт ознак моделі виявлення. Вивчені карти уваги виділяють

інформативні області, що покращують здатність виявлення, а також дозволяють ідентифікувати маніпульовані області обличчя.

9.2 Інструменти для пошуку діпфейків

Останнім часом все частіше зростає кількість фейкових відео, які вводять в оману звичайних користувачів. Однак збільшення таких відео також дає можливість створювати більш великі датасети для навчання нейронних мереж, які можуть визначати неправдиву інформацію.

Deepfake відео створюються за допомогою різних технологій машинного навчання та комп'ютерного зору. До основних підходів належать генеративно-змагальні мережі (GANs). GANs складаються з двох нейронних мереж: генератора і дискримінатора. Генератор створює підроблені зображення, а дискримінатор намагається визначити, чи є ці зображення справжніми, чи підробленими. Іншим підходом є використання автокодерів, які навчаються стискати зображення обличчя в компактні представлення, а потім відтворювати його з цього представлення. Такі моделі дозволяють змінювати обличчя на відео.

Для створення реалістичних відео важливо не лише підробити обличчя, але й синхронізувати його рухи з аудіо. Для цього використовуються технології, які відстежують рухи обличчя та накладають їх на обличчя іншої людини. В деяких випадках для створення deepfake використовуються 3D моделі обличчя, які дозволяють досягти більш реалістичних результатів, особливо в динамічних сценах. Це дає змогу забезпечити точніше відображення рухів обличчя з різних кутів зйомки та освітлення.

Одним із головних підходів до визначення deepfake на відео та фото є підрахунок та аналіз кількості кліпань очима. У нормальних умовах людська око кліпає з періодичністю 2–10 секунд, а тривалість кліпання

становить 0.1–0.4 секунди. Розробники deepfake в багатьох випадках не можуть забезпечити природний стан цього процесу. Тому використовуючи цей підхід до дослідження, можна виявити неточності, на які варто звернути увагу.

Загалом виділяють чотири основні категорії засобів зміни обличчя, що дозволяють створювати реалістичні deepfake відео. (рисунок 9.1)

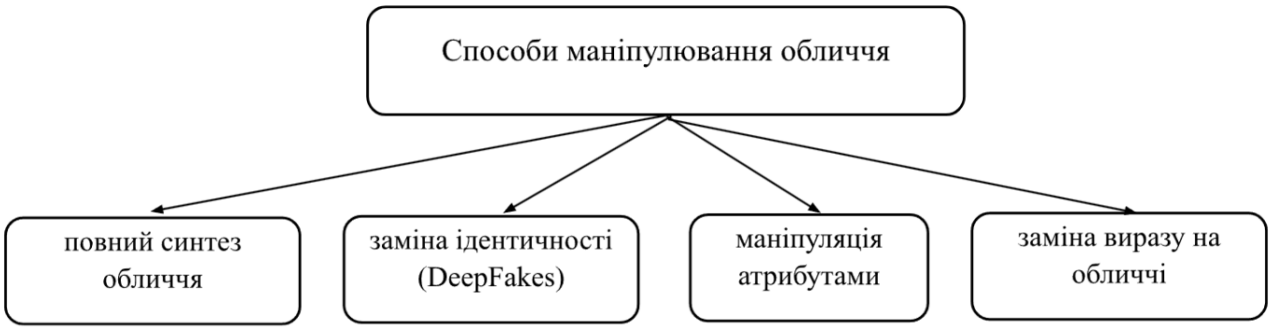


Рисунок 9.1. Способи маніпулювання обличчя

Процес створення deepfake на обличчі включає кілька етапів і використовує різні методи глибокого навчання та обробки зображень. Спочатку необхідно зібрати великий набір зображень обличчя людини, яку потрібно підробити. Ці зображення повинні включати різні вирази, кути огляду та освітлення, щоб забезпечити різноманітність даних для тренування. Зображення попередньо обробляються: обрізаються до потрібного розміру, нормалізуються за яскравістю, контрастом та іншими характеристиками, щоб підготувати їх для тренування моделі.

Зібрані зображення використовуються для тренування моделі. Один із найпоширеніших підходів — це використання автокодерів або генеративних змагальних мереж (GANs). Ці методи дозволяють створити згенероване обличчя, яке буде використовуватися для заміни оригінального на кожному кадрі відео.

Для відео необхідно відстежувати рухи обличчя у кожному кадрі. Для цього використовуються алгоритми, такі як оптичний потік, щоб визначити, як переміщається обличчя в кадрі, та застосувати ці рухи до згенерованого обличчя. Процес включає накладання згенерованого обличчя на оригінальне з урахуванням освітлення, текстур та рухів.

Якщо на відео людина розмовляє, використовуються додаткові алгоритми для синхронізації рухів губ з мовленням. Це мовлення може бути або згенероване, або попередньо записане. Якщо є звуковий супровід, наприклад, мова, то цей звук синхронізується з рухами губ на підробленому відео, щоб зробити відео більш правдоподібним.

Програмні засоби і платформи, які використовуються для створення deepfake на обличчі, наведені в таблиці 9.1.

Таблиця 9.1.

Програмні засоби і платформи, які використовуються для створення deepfake на обличчі

Назва	Технологія	Опис
DeepFaceLab	Python, GPU, opencv, tensorflow	Широкий набір інструментів для роботи з обличчями, включаючи виділення, навчання моделей та злиття облич.
Faceswap	Tensorflow, Keras, Python	Інструмент з відкритим кодом для створення deepfake. Він дозволяє замінювати обличчя на відео або фотографіях. Використовує TensorFlow і може працювати на CPU або GPU.
Zao	Android	Додаток для мобільних пристроїв, який швидко став популярним завдяки своїй здатності створювати deepfake за кілька секунд. Достатньо завантажити своє фото і додаток вставить ваше обличчя у відомі сцени з фільмів або відео.
Reface	iOS, Android	Мобільний додаток українського виробництва, що дозволяє швидко замінювати обличчя на гіфках та відео. Працює на основі штучного інтелекту та дозволяє легко створювати короткі відео з заміною обличчя.
Avatarify	Python	Інструмент, що дозволяє замінювати обличчя в реальному часі під час відеозв'язку. Працює на основі нейронних мереж і дозволяє відтворювати обличчя відомих людей або персонажів під час відеодзвінків. Може бути використаний з такими платформами, як Zoom або Skype

Продовження таблиці 9.1.

Deep Art Effects	iOS, Android	Комерційне програмне забезпечення, яке дозволяє створювати відео зі стилізацією та заміною обличчя. Підтримує різні стилі та алгоритми для модифікації відео.
Deepfakes Web	Python	Deepfakes Web – це онлайн-платформа для створення deepfake, яка дозволяє завантажувати відео і створювати deepfake за допомогою хмарних обчислень. Платформа бере на себе всі важкі обчислення, що робить її доступною для користувачів без потужного обладнання.

На рисунку 9.2. наведено приклад використання одного з підходів для виявлення фейків на основі кількості і частоти кліпань очей на відео.

Рисунок 9.2 Приклад використання одного з підходів для виявлення фейків на основі кількості і частоти кліпань очей

У даному випадку відбувається виділення очей на зображенні та ключових точок навколо очей з допомогою бібліотеки mediapipe. Завдяки цьому можна знайти відношення між окремими точками, що дозволить визначити чи у людини закриті очі чи ні. Надалі відбувається підрахунок частоти кліпань і формування рішення чи це фейк чи реальне обличчя.

Розглянемо сучасні інструменти для виявлення фейкових відео. DeepFake Detection Challenge (DFDC) Model — це одна з найуспішніших моделей для виявлення deepfake-відео, яка була розроблена як частина конкурсу, організованого Facebook у співпраці з іншими компаніями та

дослідницькими інститутами. DFDC Model використовує складні архітектури нейронних мереж, такі як згорткові нейронні мережі (CNN) і рекурентні нейронні мережі (RNN), які навчаються виявляти найменші ознаки підробки у відео, включаючи артефакти, що виникають під час обробки зображень. DFDC Model здатна виявляти підробки, створені різними типами алгоритмів deepfake, включаючи новітні методи, які можуть бути відсутні в інших моделях.

У таблиці 9.2 наведено аналіз інструментів для виявлення deepfake на обличчі.

Таблиця 9.2.

Порівняльний аналіз інструментів для виявлення deepfake на обличчі

Назва	Опис
DeepFake Detection Challenge (DFDC) Model [13]	DFDC дав змогу експертам з усього світу зібратися разом, порівняти свої моделі виявлення глибоких фейків, випробувати нові підходи та навчитися на роботі один одного. Об'єм набору даних: 124 тисячі відео Містить вісім алгоритмів модифікації обличчя Це модель, розроблена в рамках конкурсу, організованого Facebook, Microsoft, Amazon та іншими технологічними компаніями. Модель навчається на великому наборі даних, щоб виявляти різні типи deepfake, навіть коли їх створено за допомогою різних технік
FaceForensics++	FaceForensics++ є набором інструментів та даних для навчання і тестування алгоритмів розпізнавання підроблених відео. Використовується дослідниками та розробниками для створення моделей, які можуть розпізнавати deepfake на основі аналізу текстури шкіри, рухів очей та інших характеристик.
XceptionNet	XceptionNet – це вдосконалена нейронна мережа для аналізу відео та зображень, яка може розпізнавати deepfake з високою точністю. Мережа використовує згорткові шари для виявлення особливих ознак на зображеннях, які можуть вказувати на те, що обличчя було змінено.
Amber Authenticate	Amber Authenticate – це система, що використовує блокчейн для підтвердження автентичності відео. Вона створює цифрові підписи для кожного кадру відео під час запису, що дозволяє відстежувати зміни та виявляти підробки.

Продовження таблиці 9.2.

Intel FakeCatcher	FakeCatcher – це інструмент від Intel, який аналізує відео для виявлення ознак deepfake на основі аналізу руху крові через капіляри на обличчі. Це унікальний підхід, який відстежує мікроскопічні зміни кольору шкіри, пов'язані з кровообігом, які важко відтворити у deepfake.
Forensically	Це онлайн-інструмент для аналізу зображень, який може використовуватися для виявлення ознак маніпуляції, включаючи deepfake. Forensically дозволяє проводити аналіз на рівні пікселів, виявляти зміни в JPEG-компресії, аналізувати тіні та інші аспекти, які можуть вказувати на підробку.

9.3 Висновки з проведеного дослідження

На основі аналітичного підходу проведено аналіз наукових публікацій в галузі формування та визначення діпфейків на відео та зображенні, що дозволило виділити основні сучасні технології такі, як: згорткові нейронні мережі, рекуретні нейронні мережі та датасети, які використовуються для навчання нейронних мереж. Визначено підходи до способів створення фейків на відео з елементами маніпуляції людським обличчям;

Проведено аналіз сучасних програмних засобів, які використовуються для формування діпфейків, що дозволило виділити найпотужніші із них та визначити підходи за допомогою яких формуються фейкові відео. Встановлено, що Faceswap та Deepfakes Web є найпотужнішими рішеннями в даній сфері;

Проведено аналіз інструментів для виявлення deepfake на обличчі, що дозволило виділити основні технології, бібліотеки, датасети, які використовуються для пошуку фейків. DeepFake Detection Challenge (DFDC) Model, Forensically та інші володіють значним потенціалом та об’ємом вибірки для якісного визначення фейків з людським обличчям.

Список використаних джерел:

1. M. Quadir, P. Agrawal and C. Gupta, "A Comparative Analysis of Deepfake Detection Techniques: A Review," 2023 6th International Conference on

Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 1035-1041, doi: 10.1109/IC3I59117.2023.10397938.

2. Guhagarkar, N., Desai, S., Vaishyampayan, S., & Save, A.M. (2021). DEEPFAKE DETECTION TECHNIQUES: A REVIEW.

3. Akhtar, Z., Mouree, M. R., & Dasgupta, D. (2020). Utility of deep learning features for facial attributes manipulation detection. Paper presented at the 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI).

4. Wubet, W. M. (2020). The deepfake challenges and deepfake video detection. Int. J. Innov. Technol. Explor. Eng, 9(6), 789-796.

5. Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), e1520.

6. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

7. Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14923-14932).

8. ST S, Ayoobkhan MUA, V KK, Bacanin N, K V, Štěpán H, Pavel T. 2022. Deep learning model for deep fake face recognition and detection. PeerJ Computer Science 8:e881 https://doi.org/10.7717/peerj-cs.881

9. Tolosana, R., Romero-Tapiador, S., Fierrez, J., & Vera-Rodriguez, R. (2021, January). Deepfakes evolution: Analysis of facial regions and fake detection performance. In international conference on pattern recognition (pp. 442-456). Cham: Springer International Publishing.

10.Pashine, S., Mandiya, S., Gupta, P., & Sheikh, R. (2021). Deep fake detection: survey of facial manipulation detection solutions. arXiv preprint arXiv:2106.12605.

11.S., K., V., M. Image quality assessment based fake face detection. Multimed Tools Appl 82, 8691–8708 (2023). https://doi.org/10.1007/s11042-021-11493-9

12.Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition (pp. 5781-5790).

13.Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854.

14.Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

10

НЕЙРОМЕРЕЖЕВИЙ АНСАМБЛЕВИЙ МЕТОД ДО КЛАСИФІКАЦІЇ ДІПФЕЙКІВ ІЗ ВИКОРИСТАННЯМ ВИБОРУ ЗОЛОТИХ КАДРІВ

Ліп'яніна-Гончаренко Христина[1], Мельник Назар[1], Івасечко Андрій[1], Телька Микола[1], Соя Мар'яна[1],
[1]Західноукраїнський національний університет, вул. Львівська, 11, м. Тернопіль, 46000, Україна

10.1 Актуальність проблеми та огляд сучасних методів детекції діпфейків

Актуальність дослідження детекції діпфейків у сучасному світі, особливо в контексті гібридних воєн, є надзвичайно високою. Використання діпфейків для маніпуляції громадською думкою, політичними рішеннями та міжнародними відносинами вже стало реальністю, що демонструє потенційні ризики для стабільності держав і безпеки особистості. Технології діпфейків можуть використовуватися для створення фальсифікованих доказів, що спотворюють реальні події або поведінку впливових особистостей, що може спричинити міжнародні конфлікти або внутрішні непорозуміння в країнах [1-2].

На тактичному рівні, діпфейки можуть бути використані для сіяння плутанини серед військових, підриваючи моральний дух або навіть викликаючи невиправдані ретирації через візуально підтверджене, але фальсифіковане проникнення ворожих сил. Це вказує на серйозні загрози,

які діпфейки становлять для воєнних стратегій та реальної безпеки націй [3]. Наприклад, росія використовує гібридні військові тактики, що включають кібератаки та дезінформацію, щоб підірвати суспільну стабільність в Україні, демонструючи, як технології можуть бути використані як інструмент воєнізованої агресії [4-5].

В контексті глобальної політики, діпфейки використовуються для дискредитації політичних лідерів, маніпулювання громадською думкою та навіть втручання у виборчі процеси, як це було з різними випадками використання діпфейків під час виборів в США та інших країнах. Підвищення обізнаності та розвиток технологій детекції є ключовими для захисту демократичних процесів і національної безпеки від таких загроз [1].

Розвиток та впровадження регуляторних заходів, таких як Європейський цифровий акт, який запроваджує штрафи для соціальних мереж за неналежне модерування штучно створених дезінформаційних матеріалів, підкреслює глобальне визнання необхідності контролю над розповсюдженням і використанням діпфейк технологій. Ці заходи вказують на серйозність проблеми та необхідність міжнародної кооперації для боротьби з новими формами кіберзагроз [1].

Враховуючи ці виклики, у даному дослідженні ставляться такі гіпотези:

Гіпотеза 1: Використання ансамблевого нейромережевого підходу дозволить підвищити точність класифікації діпфейкових відео завдяки поєднанню сильних сторін різних архітектур нейромереж.

Гіпотеза 2: Вибір "золотих кадрів", які є найбільш інформативними для аналізу, дозволить зменшити обсяг обчислень без втрати якості класифікації.

Гіпотеза 3: Аналіз активацій моделей за допомогою Grad-CAM покаже, що різні архітектури акцентують увагу на різних ознаках, що може бути використано для покращення детекції діпфейків.

Отже, мета цього дослідження полягає в розробці передових технологій детекції діпфейків, зокрема через використання гібридного нейромережевого ансамблю, який базується на виділенні "золотих кадрів". Цей підхід покликаний суттєво підвищити точність виявлення діпфейків і забезпечити їх ефективну ідентифікацію в умовах зростаючих загроз, пов'язаних із гібридною війною та інформаційною маніпуляцією. В рамках дослідження акцентується на розробці методів, які здатні протистояти сучасним викликам, зокрема у контексті актуальних міжнародних конфліктів, таких як війна в Україні, де технології діпфейків використовуються для дезінформації та підриву стабільності суспільства. Результати даного дослідження можуть стати основою для створення ефективних рішень, які підвищать обізнаність та стійкість до інформаційних загроз у різних сферах, зокрема у медіа та соціальних мережах.

В останні роки технології DeepFake, що дозволяють маніпулювати мультимедійним контентом, зокрема відео та зображеннями, викликали значний інтерес як у наукових колах, так і в широкій громадськості через їхні потенційні ризики для особистої безпеки та інформаційної достовірності. Розробка ефективних методів детекції DeepFake стала ключовим напрямком досліджень у галузі комп’ютерного зору та штучного інтелекту. Високопродуктивне виявлення відео DeepFake з використанням мережей CNN, зорієнтованих на специфічні цільові регіони та ручне вилучення дистиляції, запропоноване Tran et al. [6], є одним з найновіших підходів у цій області. Цей метод відрізняється використанням

специфічних областей відео для підвищення точності моделей із збереженням оптимального розміру моделі.

З іншого боку, Zhang et al. [7] вказують на потребу використання гетерогенного ансамблю ознак для покращення точності детекторів DeepFake. Вони запропонували метод, що інтегрує різні ознаки, такі як сірий градієнт, спектральні ознаки та текстурні ознаки. Цей підхід демонструє кращу точність порівняно з кількома сучасними детекторами DeepFake, що підкреслює значення ансамблю ознак у виявленні маніпуляцій.

Методи, засновані на витонченій класифікації з використанням тонких, загальнозначущих ознак, як показано у дослідженні Nadimpalli та Rattani [8], забезпечують високу ефективність на різних наборах даних і техніках маніпуляції, демонструючи переваги методів, які фокусуються на дрібніших, дискримінативних ознаках. Це підкреслює важливість врахування локальних варіацій при розробці детекторів DeepFake.

Інший інноваційний підхід представлений у роботі Guan et al. [9], де вони використовують багатофункціональну ваговану модель, що базується на мета-навчанні. Ця модель покращує загальну здатність до виявлення шляхом оптимізації здатності виявляти ознаки у різних областях.

Нарешті, важливо відзначити роботу Khan et al. [10], які проводять порівняльний аналіз різних технік детекції відео DeepFake, підкреслюючи сильні та слабкі сторони кожного підходу і вносячи свій вклад у вдосконалення параметрів гіперпараметрів для покращення загальної точності і ефективності моделей.

Chakraborty та Naskar [11] досліджують вплив людської фізіології та фізичної біомеханіки обличчя на розробку надійних детекторів DeepFake. Вони аналізують як фізіологічні сигнали можуть бути використані для підвищення точності детекторів, які часто стикаються з проблемами

демографічних і соціальних упереджень. Цей підхід дозволяє знизити помилки, характерні для інших методів детекції, що спираються на загальні характеристики зображень чи відео.

В оглядовій роботі Abbas і Taeihagh [12] систематично розглядають методи детектування та генерування DeepFake з використанням штучного інтелекту. Автори досліджують ключові алгоритми, платформи та інструменти, виявляючи як ці технології можуть бути впроваджені в різних контекстах для боротьби з поширенням фальшивих новин. Вони також наголошують на практичних викликах і тенденціях у реалізації політик, що протидіють поширенню DeepFake.

Дослідження Casu et al. [13] висвітлює новий аспект у викликах, пов'язаних із детекцією DeepFake, зосереджуючись на впливі когнітивних упереджень на прийняття рішень у цифровій форензиці. Вони вводять термін "Impostor Bias", що описує систематичну тенденцію сумніватися в автентичності мультимедійного контенту, часто передбачаючи, що він генерований за допомогою штучного інтелекту. Це упередження може призвести до помилкових суджень та хибних обвинувачень, що підриває надійність та вірогідність форензичних доказів. Дослідження підкреслює, як реалістичність AI-генерованих мультимедійних продуктів може посилити вплив цього упередження, вказуючи на необхідність розробки стратегій для його запобігання та протидії.

Тим часом, Firc et al. [14] розглядають DeepFakes як загрозу для систем розпізнавання облич та голосу, надаючи детальний огляд інструментів та векторів атак у візуальних та аудіо доменах. Вони аналізують, як DeepFakes можуть впливати на біометричні системи, зокрема через спуфінг, та визначають категорії DeepFake із відповідними інструментами для їх створення, наборами даних та методами детекції. Основний внесок полягає в аналізі векторів атаки, з урахуванням різниць

між категоріями DeepFake та звітів про реальні атаки, для оцінки загроз, які вони представляють для обраних категорій біометричних систем.

Lee et al. [15] у своїй роботі "ClueCatcher" використовують новаторський підхід до виявлення DeepFake, зосереджуючись на незалежних ознаках у різних доменах. Вони ідентифікують і використовують такі ознаки, як невідповідності у кольорах облич, артефакти на межах синтезу, та відмінності у якості між обличчями та необличними регіонами. Використання мультипотокових конволюційних нейронних мереж та оцінювача міжпатчевої неподібності дозволяє цій моделі ефективно виявляти унікальні ознаки DeepFake, значно підвищуючи загальну ефективність та узагальнюваність детекції.

Naitali et al. [16] роблять огляд широкого спектру тем, пов'язаних із DeepFake, включаючи методи генерації, детекції, наявні датасети, виклики та напрямки майбутніх досліджень. Їхня робота надає фундаментальне розуміння проблем DeepFake і висвітлює потенційні ризики, які DeepFake створює у сферах, таких як дезінформація, політичні маніпуляції, репутаційний збиток та шахрайство. Автори також зосереджуються на сучасних методах детекції і виявляють потребу в подальших дослідженнях для розвитку стратегій мінімізації загроз від DeepFake.

Dincer et al. [17] представили інноваційний метод виявлення відео DeepFake, який використовує інформацію золотого перетину для вибору кадрів з відео. Цей підхід дозволяє вибирати не випадкові, а специфічні кадри, що потенційно можуть покращити продуктивність детектування шляхом зосередження на значущих регіонах обличчя. Метод використовує три різні способи екстракції ознак (VGG19, EfficientNet B0, EfficientNet B4) та дві моделі капсульних мереж (CapsuleNet та ArCapsNet), що демонструють можливості глибшого і точнішого аналізу даних. Виконані оцінки продуктивності на двох складних наборах даних для виявлення

DeepFake, Celeb-DF та DFDC-P, показали високі результати, особливо злиття кращих моделей дало високу точність і площу під кривою (AUC), зокрема 93.63% ACC і 99.14% AUC для Celeb-DF.

Alarfaj та Khan [18] детально досліджують класифікацію фейкових новин, використовуючи методи машинного і глибинного навчання. Вони використовують різні моделі машинного навчання, такі як мультиноміальний, гауссівський та Бернуллієві наївні баєсівські класифікатори, логістичну регресію, та агресивний класифікатор без пасиву. В доповненні до традиційних моделей, автори вивчають глибинні навчальні моделі, такі як LSTM та CNN-LSTM, які показали вищу точність і більш стабільну класифікацію порівняно з традиційними методами. Ці результати підкреслюють потенціал глибинних навчальних моделей для ефективного боротьби з розповсюдженням фейкових новин.

Таблиця 10.1.

Огляд найближчих аналогів

Автор(и)	Рік	Опис методу	Особливості методу	Новизна методу
Tran et al. [6]	2021	Використання CNN з орієнтацією на специфічні цільові регіони.	Висока продуктивність за рахунок ручного вилучення дистиляції.	Оптимізація розміру моделі для кращої продуктивності.
Zhang et al. [7]	2022	Гетерогенний ансамбль ознак для детекції.	Інтеграція різних ознак (сірий градієнт, спектральні ознаки та текстурні ознаки) для покращення точності.	Підвищення точності за рахунок ансамблю ознак.
Guan et al. [9]	2023	Мультифункціональна вагована модель на базі мета-навчання.	Використання RGB та частотних доменів для покращення узагальнення.	Висока узагальненість та адаптація до різних даних.
Lee et al. [15]	2023	"ClueCatcher" з відбором доменно-незалежних ознак.	Фокус на невідповідності у кольорах облич, артефактах на межах синтезу та відмінностях у якості між обличчями та необличними регіонами.	Ефективність в реальних сценаріях.

Розглянувши існуючі рішення в галузі детекції діпфейків (таблиця 10.1), можна відзначити, що наш метод, який базується на використанні

нейромережевого ансамблю з відбором золотих кадрів, виявляє суттєві переваги у порівнянні з іншими підходами. Особливо варто відмітити аналогічний метод Guan et al. [9], який також використовує складні ансамблі моделей для підвищення точності. Обидва методи демонструють високу ефективність у різноманітних умовах, проте наш підхід вирізняється кращою оптимізацією обчислювальних ресурсів та адаптивністю, завдяки цільовому відбору кадрів, що дозволяє зосередитись лише на найбільш інформативних частинах відео.

Незважаючи на значний прогрес у розробці методів детекції діпфейків, сучасні підходи мають певні обмеження.

По-перше, більшість моделей глибокого навчання, зокрема CNN-архітектури, демонструють високу точність на певних наборах даних, але суттєво втрачають ефективність при зміні джерела даних або якості відео.

По-друге, деякі підходи, такі як методи, що базуються на спектральному аналізі чи ручному виділенні ознак, є чутливими до змін освітлення та варіацій у відеопотоках.

По-третє, більшість методів опрацьовують відео покадрово, що призводить до значних обчислювальних витрат і обмежує їхню практичну застосовність у реальному часі.

Ці проблеми обґрунтовують необхідність розробки більш ефективних методів, здатних зберігати високу точність при меншій витраті ресурсів, що і є основною метою цього дослідження.

Відповідно, новизна нашого методу полягає у здатності ефективно використовувати золоті кадри для тренування моделей, що суттєво підвищує швидкість обробки та точність результатів у порівнянні з традиційними методами, які аналізують кожен кадр відео. Цей підхід не лише підвищує ефективність але й забезпечує більшу узагальнюваність у

виявленні діпфейків у різноманітних сценаріях, що робить його ідеально підходящим для застосувань у медіа та соціальних мережах.

10.2 Розробка та реалізація ансамблевого методу

Для підвищення точності розпізнавання діпфейків запропоновано нейромережевий ансамблевий метод (Рисунок 10.1), що базується на виділенні "золотих кадрів" — найбільш інформативних кадрів, які відображають суттєві зміни у сцені. Цей підхід дозволяє значно скоротити обсяг даних, зосереджуючи увагу на ключових моментах, що забезпечує більш точне навчання моделей глибокого навчання та підвищує ефективність кінцевого прогнозу. На рисунку 1 зображено процес виділення золотих кадрів із відео та подальшого ансамблевого навчання, який дозволяє досягти вищої точності класифікації діпфейк-контенту. Далі представлено кроки запропонованого методу:

Крок 1: Збір даних. Початковий етап обробки відео включає збір даних з відповідних джерел та їх організацію. Відеодані зберігаються у вигляді файлів, кожен з яких має метадані, що містять інформацію про клас (наприклад, "FAKE" чи "REAL"). Ці метадані використовуються для позначення міток класу під час підготовки тренувальних і валідаційних вибірок. Нехай M позначає набір метаданих, що містить інформацію про кожен відеофайл та його клас, а V — директорія з відеофайлами.

Крок 2: Функція для відбору золотих кадрів з відео [17]. Ця функція ідентифікує та вибирає найбільш інформативні кадри з відео, які визначаються на основі значних змін у сцені. Процес вибору кадрів здійснюється шляхом порівняння кожного кадру з попереднім, з використанням абсолютної різниці, перетвореної у відтінки сірого, щоб обчислити загальну інтенсивність зміни сцени. Математично це можна представити як обчислення середньої суми інтенсивності різниці між двома

послідовними кадрами, де, якщо вона перевищує заданий поріг, кадр додається до вибірки золотих кадрів. Цей процес продовжується до тих пір, поки не буде вибрано задану кількість кадрів або до кінця відео, забезпечуючи вибірку найбільш значущих сцен. Нехай V позначає відео, яке складається з N кадрів. Функція вибирає набір кадрів F, які значно відрізняються один від одного за заданим порогом інтенсивності зміни сцени θ.

2.1. Ініціалізація. Ініціалізуємо набір F з першого кадру f_0, який ми зчитуємо з V і додаємо до F. Відповідно позначаємо f_0 як $fprev$.

2.2. Ітераційний процес. Проводимо ітерацію крізь кожен $i-$ й кадр V, де і змінюється від Δi до N з кроком $\Delta i = \max\left(1, \left\lfloor \frac{N}{num_{frames}} \right\rfloor\right)$:

Зчитуємо $i-$ й кадр f_i з V.

Обчислюємо абсолютну різницю D між $fprev$ та f_i:

$D = cv2.cvtColor\left(cv2.absdiff\left(f_{prev}, f_i\right), cv2.COLOR_{BGR2GRAY}\right)$

Розраховуємо метрику різниці S як середнє значення пікселів у D:

$$S = \frac{\sum D}{m \times n}$$

де $m \times n$— розмір кадру після масштабування.

Якщо $S > \theta$, додавання f_i до F і оновлення $f_{prev} = f_i$.

2.3. Умова завершення. Процес ітерації завершується, коли кількість кадрів у F досягне num_{frames} або коли усі кадри V були перевірені.

Крок 3: Попередня обробка даних з використанням золотих кадрів полягає в тому, що функція спочатку перевіряє наявність кожного відеофайлу, вказаного у метаданих, після чого витягує з кожного відео задану кількість золотих кадрів, які відрізняються значними змінами сцен. Математично процес можна описати як перебір метаданих M, що містить

пари назв файлів та додаткових даних, з наступним викликом функції $F(f_i, k)$, що витягує k золотих кадрів для кожного відео.

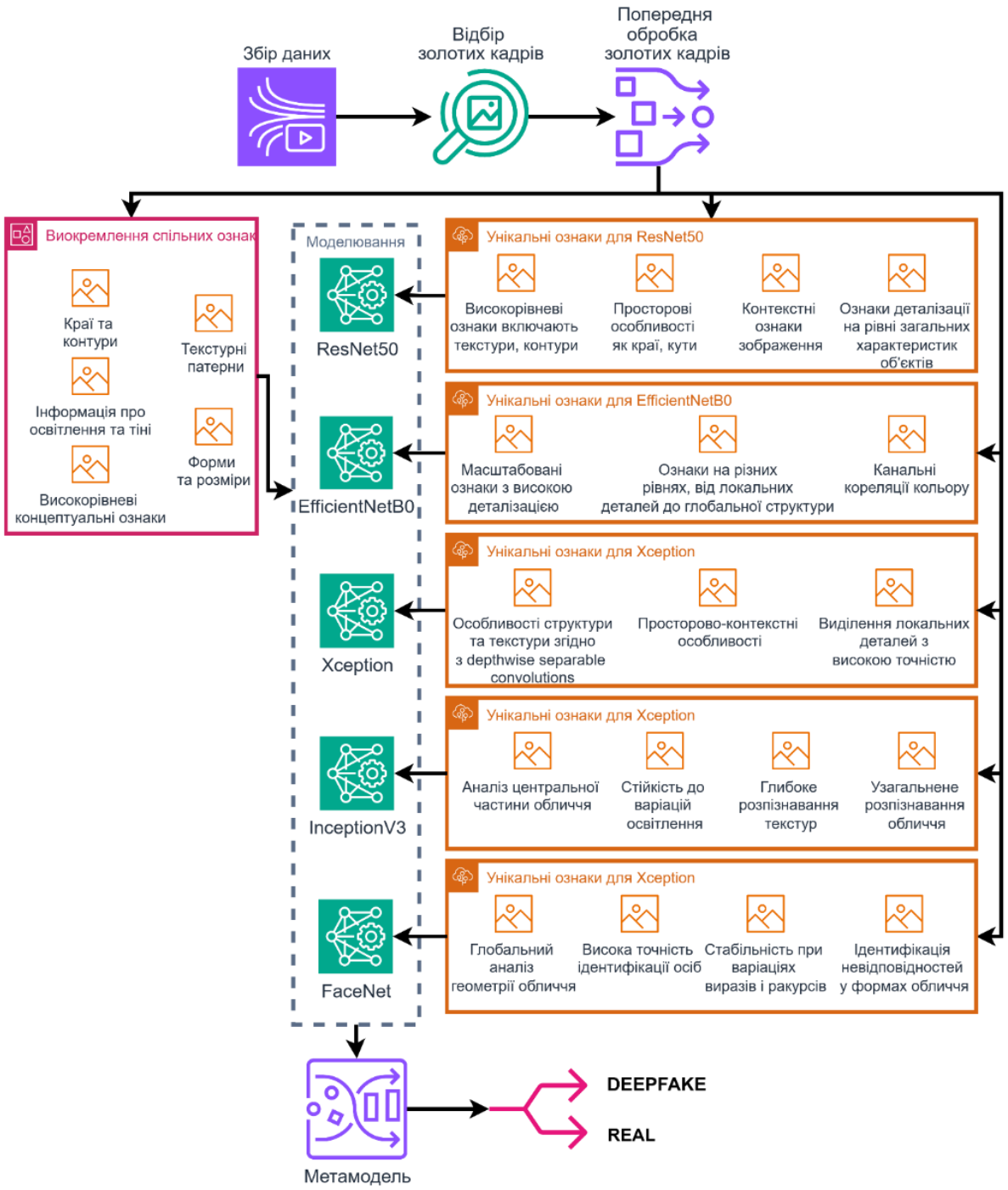


Рисунок 10.1. Структура нейромережевого ансамблевого методу до класифікації діпфейків із використанням вибору золотих кадрів

Нехай M позначає метадані для відеофайлів, де кожен елемент $M_i = (f_i, d_i)$, де f_i — ім'я файлу, а d_i — додаткові дані (такі як мітки класу). Набір

V позначає директорію, де зберігаються відеофайли. Нехай $F(f_i, k)$ позначає функцію вибору золотих кадрів для відеофайлу f_i з максимальною кількістю k кадрів.

3.1. Попередня обробка. Ініціалізація порожніх списків X та Y для збереження оброблених кадрів та їх міток відповідно. Для кожного елемента (f_i, d_i)у M, якщо файл існує: $os.path.exists\big(os.path.join(V, f_i)\big)$:

Визначення $G_i = F(f_i, k)$ — золотих кадрів з файлу f_i.

Додавання оброблених кадрів до X та міток до Y на основі мітки класу $d_i['label']$: $y_i = 1$ якщо $d_i['label'] = 'FAKE'$, інакше $y_i = 0$.

3.2. Повернення результатів. Конвертація списків X та Y у масиви NumPy: $np.array(X), np.array(Y)$.

Крок 4: Завантаження та попередня обробка навчальних даних, а також їх поділ на тренувальні та валідаційні вибірки може бути описано у наступному математичному вигляді:

4.1. Завантаження та попередня обробка даних. Використовуємо функцію $preprocess_{video_{data(M,V)}}$, де M - метадані, а V - шлях до директорії з відео. Функція повертає набори X та Y, які містять оброблені золоті кадри та відповідні мітки (0 або 1 залежно від класу справжності відео).

4.2. Поділ даних на тренувальні та валідаційні вибірки. Застосування функції $traintest_{split}(X, Y, test_{size} = 0.2, random_{state} = 42)$для розділення даних на тренувальний набір (X_{train}, Y_{train}) та валідаційний набір (X_{val}, Y_{val}), де параметр $test_{size} = 0.2$ вказує, що 20% даних буде виділено для валідації, а решта 80% - для тренування. Параметр $random_{state} = 42$ забезпечує відтворюваність розподілу.

Крок 5: Створення базових моделей на основі ResNet50 [19], EfficientNetB0 [20] , Xception [21], InceptionV3 [22] та Facenet [23], що

базується на основі InceptionResNetV2 [24], можна описати наступним чином: для кожної моделі використовується попередньо навчена архітектура, яка виконує початкову обробку зображень через згорткові шари для витягування ознак (Таблиця 1). Кожна модель [25] має структуру, що включає послідовність обробки, яка складається з попередньо навчених згорткових шарів, повнозв'язних шарів з активацією Swish, шарів Dropout для регуляризації, та виходу, що використовує активацію Sigmoid для прогнозування бінарних класів. Отже, може бути описано у наступному математичному вигляді:

5.1. Визначення вхідних даних. Нехай $X \in R^{N\times H\times W\times C}$ представляє вхідний тензор, де N — кількість зразків, $H = 224, W = 224$, і $C = 3$ — розміри зображення (висота, ширина, кількість каналів).

5.2. Використання попередньо натренованих базових моделей. Для кожної моделі використовується попередньо натренована архітектура без верхніх шарів (тобто без шарів класифікації):

$$ResNet50 \text{ з } LSTM\text{: } FResNet(X) = ResNet50(X)$$

$$EfficientNetB0\text{: } FEffNet(X) = EfficientNetB0(X)$$

$$Xception\text{: } FXception(X) = Xception(X)\ F$$

$$InceptionV3\text{: } FInception(X) = InceptionV3(X)$$

$$Facenet\text{: } FFacenet(X) = InceptionResNetV2(X)$$

5.3. Flattening виходу базової моделі. Отриманий після базової моделі тензор ознак FModel(X) перетворюється у плаский вектор:

$$Z = Flatten(FModel(X)) \in Rd$$

де d — розмірність плаского вектора ознак, яка залежить від архітектури базової моделі.

5.4. Повнозв'язний шар з активацією Swish. До плаского вектора застосовується повнозв'язний шар з 512 нейронами та функцією активації Swish:

$$Z' = Swish(W1Z + b1)$$

де: $W_1 \in R512 \times d$ — матриця ваг; $b_1 \in R512$ — вектор зміщення.

Функція Swish визначається як $Swish(x) = x \cdot \sigma(x)$, де $\sigma(x)$— сигмоїдна функція.

5.5. Шар Dropout для регуляризації. Для запобігання перенавчанню застосовується шар Dropout з ймовірністю занулення нейронів p=0.5:

$$Z'' = Dropout(Z', p)$$

5.6. Додаткові шари для ResNet50 з LSTM. Тільки для моделі ResNet50 з LSTM виконується наступне:

Розширення виміру:

$$Z''' = ExpandDims(Z'', axis = 1)$$

Це перетворює вектор Z'' у тензор для сумісності з LSTM-шаром.

LSTM-шар:

$$ZLSTM = LSTM(256)(Z''')$$

де LSTM має 256 одиниць пам'яті.

Відмінні властивості: Використання LSTM дозволяє моделі враховувати послідовні або часові залежності у вхідних даних.

Крок 7: Вихідний шар з активацією Sigmoid

Застосовується вихідний повнозв'язний шар з одним нейроном та сигмоїдною активацією для бінарної класифікації:

Для моделей без LSTM:

$$\hat{y} = \sigma(W_2 Z'' + b_2)$$

Для ResNet50 з LSTM:

$$\hat{y} = \sigma(W_2 Z_{LSTM} + b_2)$$

де:

$W_2 \in R^{1 \times n}, n = 512$ для моделей без LSTM та n=256 для моделі з LSTM.

$b_2 \in R$ — зміщення.

$\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоїдна функція.

Таблиця 10.2.

Порівняння ознак, витягнутих моделями ResNet50, EfficientNetB0, Xception, InceptionV3 та Facenet

Тип ознаки	ResNet50	EfficientNetB0	Xception	InceptionV3	Facenet
Краї та контури об'єктів	Виявлення меж об'єктів.	Виявлення меж з оптимізацією обчислювальних ресурсів.	Виявлення контурів з глибокою деталізацією.	Виявлення країв через багатоканальні операції для загальних об'єктів.	Точне виявлення країв облич для розпізнавання особливостей.
Текстурні патерни	Визначення текстур поверхонь.	Текстурні деталі з різних масштабів.	Визначення комплексних текстур.	Широкий аналіз текстур, адаптований для загальних об'єктів.	Витяг текстур, спеціально налаштований для ознак облич.
Форми та розміри	Розпізнавання базових форм і розмірів.	Ефективне розпізнавання форм на багатьох рівнях.	Визначення складних форм і геометричних особливостей.	Узагальнене виявлення форм для різноманітних об'єктів.	Детальне розпізнавання форм обличчя для точної ідентифікації.
Освітлення та тіні	Аналіз змін освітлення.	Балансування локальних і глобальних змін освітлення.	Висока чутливість до тіней і освітлення.	Стійкість до змін освітлення завдяки мультискейлінгу.	Аналіз тонких варіацій освітлення, специфічний для обличчя.
Концептуальні ознаки	Розпізнавання облич, автомобілів тощо.	Розпізнавання об'єктів на основі спрощених ознак.	Висока ефективність у розпізнаванні складних об'єктів.	Узагальнене розпізнавання об'єктів з різних категорій.	Оптимізоване для розпізнавання облич та ідентифікації осіб.
Глибинні ознаки	Складні патерни (форми, текстури, контури).	Складні патерни на різних масштабах.	Структурні та текстурні особливості.	Узагальнені ознаки на глибоких рівнях, ефективні для широкого спектра об'єктів.	Глибинні риси облич для диференціації навіть схожих осіб.
Просторові особливості	Геометричні деталі (прямі лінії, кути).	Локальні деталі та глобальна структура.	Складні взаємодії між елементами.	Геометричні патерни, адаптовані для різних масштабів об'єктів.	Просторові особливості для точного визначення деталей облич.
Кольорові особливості	-	Канальні кореляції кольору.	Окреме виділення просторових та канальних.	Чутливість до кольорових каналів для загальних об'єктів.	Врахування тонких кольорових відтінків, важливих для обличчя.

Продовження таблиці 10.2.

Контекстні особливості	Взаємодія об'єктів у сцені.	Локальні та глобальні зв'язки між об'єктами.	Взаємодії між об'єктами в сцені.	Виявлення зв'язків між об'єктами в складних сценах.	Визначення контексту облич у сцені для підвищення точності.
Ознаки на різних рівнях	-	Баланс між деталями та загальними патернами.	Локальні деталі та загальні характеристики.	Різні рівні ознак для різних категорій об'єктів.	Локальні та загальні ознаки для ідентифікації облич.
Ознаки деталізації	Загальні характеристики об'єктів.	Висока деталізація дрібних об'єктів.	Точне виділення деталей та текстур.	Різні рівні деталізації для широкого спектра об'єктів.	Супердеталізація рис облич для точного розпізнавання.
Особливості локальної структури	Фокус на загальних формах та текстурах.	Баланс між локальними та глобальними ознаками.	Фокус на локальних деталях.	Локальні та глобальні ознаки для точного аналізу сцен.	Фокус на локальних особливостях облич для надійної ідентифікації.

Крок 6: Тренування базових моделей на основі ResNet50, EfficientNetB0, Xception, InceptionV3 та Facenet можна математично представити так:

6.1. Компіляція моделей. Нехай $M_{ResNet}, M_{EfficientNet}, M_{Xception}, M_{InceptionV3}, M_{Facenet}$ — це моделі на основі ResNet50, EfficientNetB0, Xception, InceptionV3 та Facenet відповідно. Для кожної з цих моделей проводиться оптимізація з використанням алгоритму Adam, а функція втрат визначається як бінарна крос-ентропія:

$$L_{binary}(y, \hat{y}) = -(ylog(\hat{y}) + (1 - y)\log(1 - \hat{y}))$$

де y — це справжня мітка, а $\hat{y}$ — передбачення моделі. Ціль полягає у мінімізації L_{binary} з використанням оптимізатора Adam:

$$\theta_{new} = \theta_{old} - \eta \nabla_{\theta} L_{binary}$$

де θ — це параметри моделі, а η — це швидкість навчання.

6.2. Тренування моделей. Для кожної моделі M з тренувальними даними (X_{train}, Y_{train}) і валідаційними даними (X_{val}, Y_{val}) процес тренування моделі можна описати як мінімізацію функції втрат L_{binary}

через 10 епох з використанням розміру пакету (batch size) 8. Це можна записати наступним чином:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L_{binary}\big(y_i, M(X_i;\theta)\big)$$

де N — це кількість зразків у тренувальному наборі, X_i— вхідні дані, y_i — мітка для $i-$го зразка, а θ — параметри моделі.

Процес тренування виконується для кожної моделі, де на кожній епосі відбувається оновлення параметрів на основі похідної функції втрат і оцінюється точність на валідаційних даних після кожної епохи.

Крок 7: Формування мета-ознаків (прогнозів базових моделей) відбувається наступним чином:

7.1. Прогнозування для тренувальної вибірки. Для кожної базової моделі $M_{ResNet}, M_{EfficientNet}, M_{Xception}, M_{InceptionV3}, M_{Facenet}$, обчислюється передбачення для тренувальних даних X_{train}. Нехай $\hat{y}_{ResNet}(X_{train}), \hat{y}_{EfficientNet}(X_{train}), \hat{y}_{Xception}(X_{train}), \hat{y}_{InceptionV3}(X_{train}), \hat{y}_{\text{Facenet}}(X_{train})$ представляють передбачені значення базових моделей для тренувальних даних:

$$\hat{y}_{ResNet} = M_{ResNet}(X_{train}),$$
$$\hat{y}_{EfficientNet} = M_{EfficientNet}(X_{train}),$$
$$\hat{y}_{Xception} = M_{Xception}(X_{train}),$$
$$\hat{y}_{InceptionV3} = M_{InceptionV3}(X_{train}),$$
$$\hat{y}_{Facenet} = M_{Facenet}(X_{train})$$

7.2. Прогнозування для валідаційної вибірки. Аналогічно, для валідаційної вибірки X_{val} отримуємо передбачення:

$\hat{y}_{ResNet}(X_{val}), \hat{y}_{EfficientNet}(X_{val}), \hat{y}_{Xception}(X_{val}), \hat{y}_{InceptionV3}(X_{val}), \hat{y}_{Facenet}(X_{val})$

— це передбачені значення базових моделей для валідаційних даних.

7.3. Формування вхідних даних для метамоделі. Вхідні дані для метамоделі (стекування) формуються шляхом горизонтального об'єднання (конкатенації) передбачень базових моделей для тренувальної та валідаційної вибірок:

$$X_{meta}^{train} = hstack\begin{pmatrix}\hat{y}_{ResNet}(X_{train}), \hat{y}_{EfficientNet}(X_{train}), \hat{y}_{Xception}(X_{train}), \hat{y}_{InceptionV3}(X_{train}), \\ \hat{y}_{\text{Facenet}}(X_{train})\end{pmatrix}$$

$$X_{meta}^{val} = hstack\begin{pmatrix}\hat{y}_{ResNet}(X_{val}), \hat{y}_{EfficientNet}(X_{val}), \hat{y}_{Xception}(X_{val}), \\ \hat{y}_{InceptionV3}(X_{val}), \hat{y}_{Facenet}(X_{val})\end{pmatrix}$$

Ці нові матриці ознак X_{meta}^{train} і X_{meta}^{val} використовуються як вхідні дані для метамоделі, яка буде тренуватися на прогнозах базових моделей для остаточного прогнозування.

Крок 8: Навчання метамоделі полягає в тренуванні її на основі прогнозів базових моделей, використовуючи стековані ознаки, що формуються з передбачень для кожного зразка. Математично це виглядає так:

8.1. Створення метамоделі. Нехай M_{meta} — метамодель, що тренується на ознаках, отриманих з передбачень базових моделей. Ці ознаки формуються шляхом стекування прогнозів $\hat{y}_{ResNet}$, $\hat{y}_{EfficientNet}$, $\hat{y}_{Xception}$, $\hat{y}_{InceptionV3}$, $\hat{y}_{Facenet}$ для кожної вибірки:

$$X_{meta} = hstack(\hat{y}_{ResNet}, \hat{y}_{EfficientNet}, \hat{y}_{Xception}, \hat{y}_{InceptionV3}, \hat{y}_{Facenet})$$

8.2. Навчання метамоделі. Метамодель навчається на тренувальних мета-ознаках X_{meta}^{train} з відповідними мітками y_{train} , мінімізуючи функцію втрат $L(y, \hat{y})$, де $\hat{y}$ — це прогнози метамоделі:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, M_{meta}\left(X_{meta}^{train}; \theta_{meta}\right)\right)$$

8.3. Прогнозування для валідаційної вибірки. Метамодель робить прогнози на валідаційному наборі ознак X_{meta}^{val}, отримуючи $\hat{y}_{meta}(X_{meta}^{val})$.

8.4. Оцінка точності метамоделі. Точність метамоделі оцінюється через порівняння $\hat{y}_{meta}$ з мітками y_{val} :

$$Accuracy = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \mathbb{I}\,(y_i = \hat{y}_i)$$

8.5. Формування і візуалізація матриці точності. Матриця точності визначається як:

$$C = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

де TP, FP, FN, TN — кількість істинно позитивних, хибно позитивних, хибно негативних та істинно негативних відповідей відповідно, що візуалізується за допомогою теплової карти.

10.3 Реалізація

10.3.1 Case Study

Для реалізації запропонованого методу використано набір даних з конкурсу Deepfake Detection Challenge на платформі Kaggle (Kaggle, 2020) [26], найбільшого на сьогоднішній день набору даних для виявлення DeepFake. Він містить понад 100,000 відеофрагментів, які були створені з участю 3,426 акторів, використовуючи кілька методів генерації DeepFake, зокрема GAN-базовані методи. Усі відео були сгенеровані з повною згодою учасників, що зробило цей набір даних унікальним у своєму роді.

У навчальному наборі даних є 400 відео, з яких приблизно 209 оригінальних відео використовувалися для створення 400 відео. З них 323 відео мічені як "FAKE" і 77 як "REAL", що відповідає приблизно 80% фальшивих відео (Рисунок 10.2). Це підтверджує, що набір даних має велику кількість модифікованих відео порівняно зі справжніми. Кілька оригінальних відео використовувалися багаторазово для створення

фальшивих відео, з деякими оригіналами, які генерують до шести фальшивих відео.

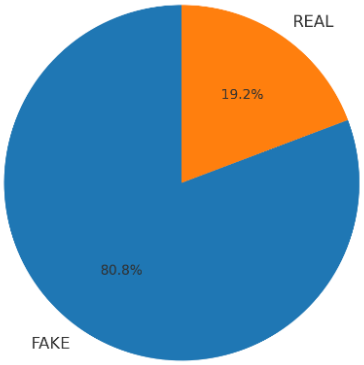


Рисунок 10.2. Розподіл міток у навчальному наборі даних

Далі використовується метод вибору золотих кадрів для зменшення обсягу оброблюваних відеоданих та виділення найбільш інформативних моментів у відеозаписах. Метод базується на аналізі змін між послідовними кадрами: з кожного відео вилучаються кадри з фіксованим інтервалом, після чого обчислюється середня різниця яскравості між поточним і попереднім кадром за допомогою абсолютної різниці пікселів у градаціях сірого. Кадр вважається суттєво відмінним, якщо значення цієї різниці перевищує встановлений поріг (30.0), що вказує на значну зміну в сцені. Наприклад (Рисунок 10.3), кадри з фейкових відео, таких як eekozbeafq.mp4 або dcuiiorugd.mp4, можуть демонструвати більше артефактів або зміни в освітленні, у порівнянні з реальними відео, такими як avmjormvsx.mp4.

Попередня обробка даних включає зміну розміру обраних кадрів до стандартних 224x224 пікселів, щоб вони відповідали вимогам вхідних шарів моделей глибокого навчання, таких як ResNet, EfficientNet та Xception. Після цього проводиться нормалізація піксельних значень до діапазону [0, 1], що сприяє стабільності та ефективності навчання моделей. Кадри та їхні відповідні мітки (0 для реальних відео, 1 для фейкових) організовуються у тривимірні тензори з формою (num_videos, num_frames, 224, 224, 3) для вхідних даних X та (num_videos,) для міток y, що забезпечує

коректне представлення відеоданих для навчання моделей. Після підготовки даних вони поділяються на тренувальну та валідаційну вибірки у пропорції 80% для навчання та 20% для валідації.

Рисунок 10.3. Вибрані золоті кадри з фейкових та реальних відео

Для аналізу послідовностей відеокадрів була розроблена архітектура, що комбінує попередньо навчені моделі ResNet50, EfficientNetB0 та Xception з LSTM для обробки тимчасових залежностей у послідовності кадрів. Вхідні дані складаються з 10 кадрів розміром 224x224 пікселів кожен, що дозволяє моделям враховувати просторові та тимчасові особливості відео. Обробка кожного кадру виконується окремо за допомогою шару TimeDistributed, після чого застосовується LSTM із 256 нейронами для моделювання залежностей між кадрами. Кожна модель тренувалася протягом 10 епох з використанням batch_size=4 та функції втрат binary_crossentropy, що забезпечує класифікацію відео на фейкові та реальні. Це поєднання дозволяє врахувати складні просторово-часові взаємодії у відео для точнішого розпізнавання маніпуляцій.

Графік (рисунок 10.4) точності показує, що серед розглянутих моделей найкращі результати на тренувальному наборі демонструє FaceNet, досягаючи 90.95% на 9-й епосі. ResNet50 також має високу точність, досягаючи 91.38%, проте її валідаційна точність зупиняється на 65.28%, що свідчить про певне перенавчання. EfficientNet демонструє стабільність із тренувальною точністю 88.10% та валідаційною 59.72%, що свідчить про гарне узагальнення моделі. Xception та Inception мали нижчу стабільність, де Xception показує різницю між тренувальною та валідаційною точністю (89.06% і 58.33% відповідно), що свідчить про труднощі узагальнення.

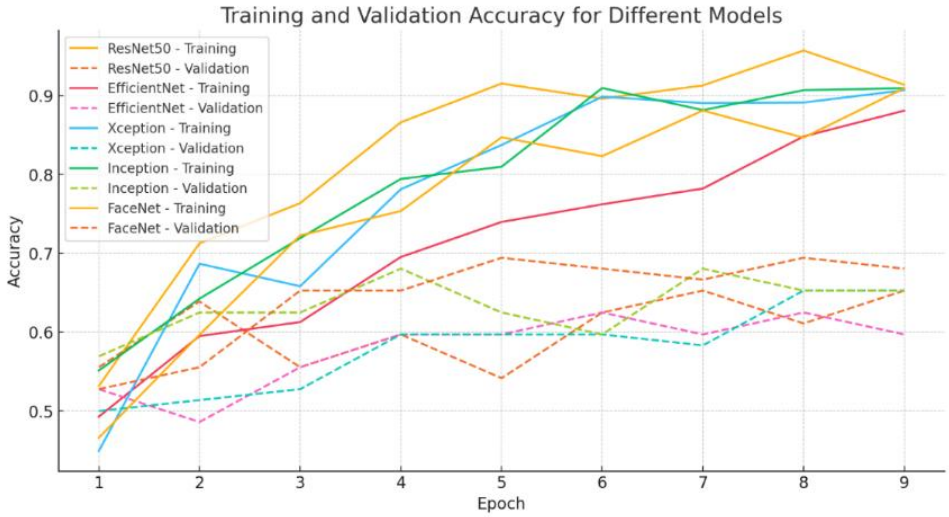


Рисунок 10.4. Зміна точності на тренувальній та валідаційній вибірках для різних моделей

Аналіз графіка (рисунок 10.5) втрат показує, що всі моделі демонструють зменшення тренувальних втрат, проте Xception і Inception мають вищі валідаційні втрати (0.9211 і 0.7736 відповідно), що вказує на ймовірне перенавчання. ResNet50 та FaceNet демонструють більш стабільне зменшення втрат із мінімальними значеннями 0.2488 і 0.2672 відповідно на тренувальних вибірках. EfficientNet показує найменшу різницю між тренувальними та валідаційними втратами, що підтверджує її стійкість до перенавчання, але при цьому має нижчу остаточну точність. Це

свідчить про те, що використання ансамблевого підходу з метамоделлю може покращити узагальнюваність моделей та підвищити їх ефективність у реальних застосуваннях.

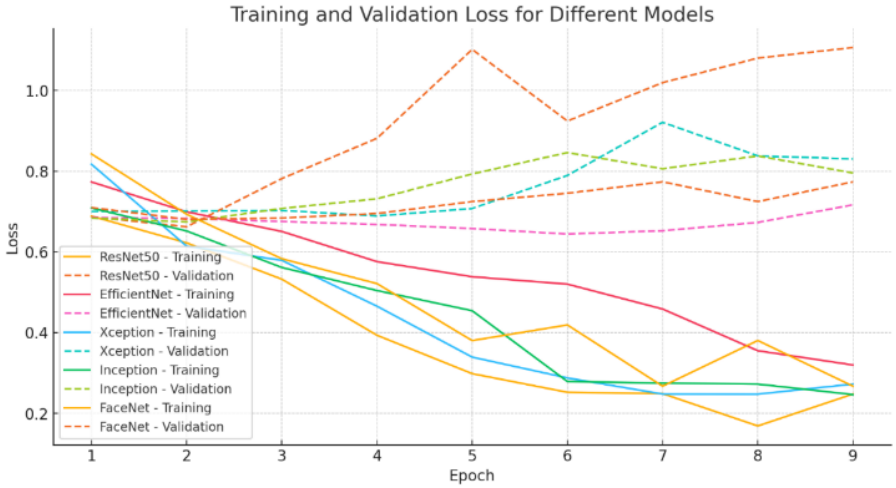


Рисунок 10.5. Динаміка втрат на тренувальній та валідаційній вибірках для різних моделей

10.3.2 Вплив ознак

Для аналізу важливих ознак, що впливають на класифікацію діпфейкових відео, було використано метод Grad-CAM, який дозволяє візуалізувати активації нейронних мереж (рисунок 10.6). Отримані теплові карти показали, що різні архітектури нейромереж акцентують увагу на різних областях обличчя (Таблиця 3). ResNet50 орієнтується переважно на текстурні особливості шкіри, зокрема на лобі та щоках, що вказує на її чутливість до глобальних спотворень. EfficientNetB0 приділяє більше уваги контурам обличчя, зокрема лініям рота та очей, що робить її ефективною у виявленні структурних аномалій. Модель Xception фокусується на областях очей і губ, що є ключовими зонами для ідентифікації діпфейків, оскільки вони часто містять неприродні артефакти, пов’язані з рухом та текстурою. Аналіз отриманих активацій свідчить про те, що комбінування різних архітектур може підвищити ефективність розпізнавання фальсифікацій за рахунок використання різних підходів до обробки візуальної інформації.

Рисунок 10.6. Візуалізація впливу ознак на класифікацію діпфейків за допомогою Grad-CAM

Таблиця 10.3.

Аналіз активацій моделей Grad-CAM у процесі класифікації діпфейків

Модель	Основні зони активації	Основний тип аномалій
ResNet50	Лоб, щоки	Текстурні спотворення, нерівномірність освітлення
EfficientNetB0	Контури рота, очей	Деформації контуру обличчя, неприродні переходи світла
Xception	Область очей, губи	Аномалії в русі, неприродні деталі текстури
InceptionV3	Центральна частина обличчя	Нестабільність кольорів, розмиття деталей
FaceNet	Вся структура обличчя	Глобальні спотворення форми та пропорцій
ResNet50	Лоб, щоки	Текстурні спотворення, нерівномірність освітлення

Отримані результати підтверджують, що різні архітектури нейронних мереж використовують відмінні підходи до виявлення ознак, характерних для діпфейків. Використання Grad-CAM дозволяє не лише інтерпретувати рішення моделей, а й ідентифікувати ключові області, що впливають на класифікацію, такі як текстурні особливості, контури обличчя та рухові

артефакти. Це відкриває можливості для подальшого вдосконалення методів розпізнавання діпфейків шляхом комбінування моделей із різними зонами чутливості для досягнення більшої точності та надійності.

10.3.3 Оцінки точності

Аналіз важливості ознак у метамоделі демонструє (рисунок 10.7), що найбільш значущим фактором у процесі класифікації діпфейків є середнє значення прогнозів моделі Xception, яке має найвищу вагу (5.1123). Другим за важливістю є медіанне значення вихідних прогнозів ResNet (4.8745), що підтверджує стабільність і значимість цієї архітектури в ансамблевому підході. Дещо нижчу, але все ще суттєву важливість мають середні значення ResNet (1.5623) та стандартне відхилення EfficientNet (1.1345), що свідчить про їхню роль у виявленні локальних відмінностей у відео. Ознаки, пов'язані з моделлю Inception (std = 0.2678, median = 0.1452), а також середнє значення Facenet (0.0817), мали значно нижчу вагу, що вказує на їхню меншу інформативність у процесі класифікації.

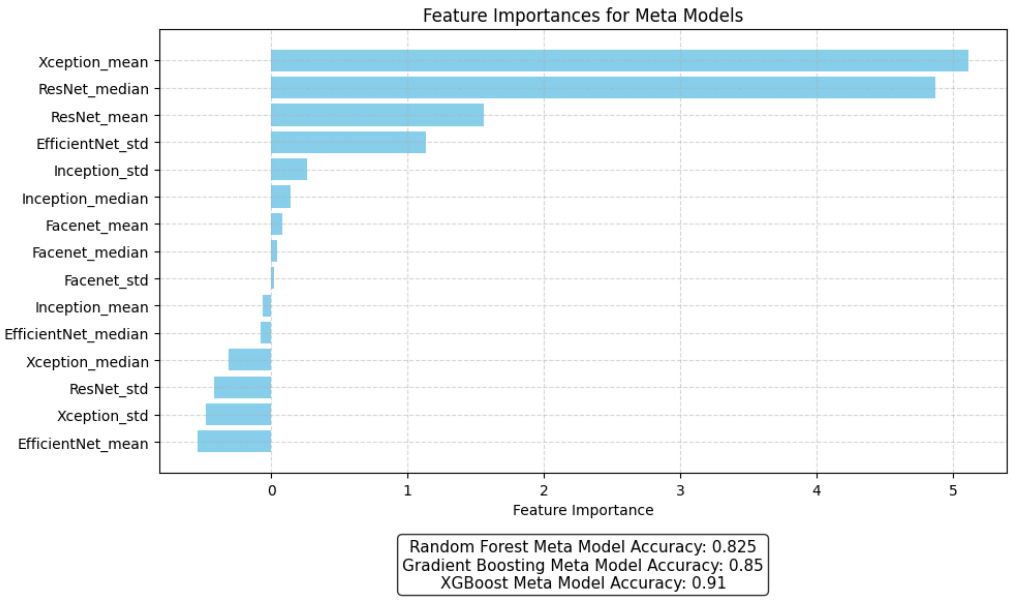


Рисунок 10.7. Важливість ознак у метамоделі для класифікації діпфейків

Оцінка точності різних метамоделей підтверджує перевагу XGBoost, який досяг максимальної точності 0.91, що значно перевищує результати Gradient Boosting (0.85) та Random Forest (0.825). Це свідчить про ефективність XGBoost у комбінуванні прогнозів базових моделей, дозволяючи досягти більш точного узагальнення. Цікаво, що деякі ознаки, такі як EfficientNet_mean (-0.5352), Xception_std (-0.4789) та ResNet_std (-0.4156), мали негативні значення важливості, що може вказувати на їхній потенційний вплив у бік зниження ефективності класифікації.

Отримані результати підтверджують ефективність використання ансамблевих моделей для детекції діпфейків і вказують на доцільність подальшої оптимізації вагових коефіцієнтів ознак у метамоделі.

10.3.4 Опис модуля та його роботи

Модуль TruScanAI (рисунок 10.8) розроблений для автоматизованого аналізу відеофайлів з метою виявлення діпфейків. Інтерфейс користувача передбачає можливість завантаження відео, після чого система здійснює його обробку із застосуванням запропонованого ансамблевого нейромережевого підходу до класифікації діпфейків із використанням методики вибору золотих кадрів. У процесі аналізу модуль автоматично екстрагує ключові кадри, аналізує текстурні, контурні та кінематичні особливості облич. Після завершення перевірки система надає користувачеві результат у вигляді класифікації відео як REAL або FAKE, доповнений візуалізацією аналізу та ключовими індикаторами, що вплинули на рішення моделі.

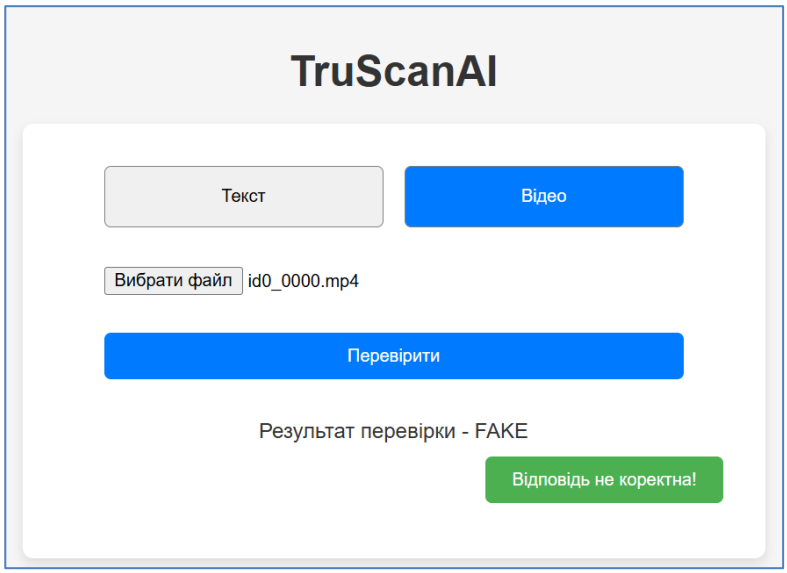


Рисунок 10.8. Інтерфейс перевірки відео на діпфейк у TruScanAI

Такий підхід забезпечує ефективну та доступну перевірку автентичності відеоконтенту як для звичайних користувачів, так і для експертів у сфері цифрової безпеки.

10.4 Обговорення

Аналіз отриманих результатів. Отримані результати підтверджують ефективність запропонованого ансамблевого підходу до класифікації діпфейків, зокрема завдяки інтеграції прогнозів ResNet50, EfficientNetB0 та Xception у метамодель XGBoost, яка продемонструвала найвищу точність серед усіх протестованих моделей (91%). Таким чином, використання методу вибору золотих кадрів у поєднанні з ансамблевим навчанням забезпечує баланс між продуктивністю та узагальненням, дозволяючи покращити детекцію діпфейків при оптимізації обчислювальних витрат.

Інтерпретація активацій нейромереж. Аналіз активацій моделей за допомогою Grad-CAM підтвердив, що різні архітектури нейромереж мають специфічні зони уваги при класифікації діпфейків. ResNet50 фокусується переважно на текстурних особливостях шкіри, зокрема на лобі та щоках, що пояснює її ефективність у виявленні глобальних артефактів.

EfficientNetB0 орієнтується на контури рота й очей, що забезпечує точніше виявлення структурних аномалій. Xception приділяє найбільшу увагу областям губ та очей, що є ключовими зонами для виявлення аномалій у русі та текстурі. Поєднання цих моделей у ансамблевому підході дозволяє компенсувати їхні індивідуальні недоліки та підвищити точність розпізнавання фальсифікацій.

Обмеження та потенційні виклики. Незважаючи на високі показники точності, запропонований метод має певні обмеження. Одним із викликів є чутливість моделей до якості вхідних відео: при низькій роздільній здатності або значному стисненні відеофайлів точність класифікації може знижуватися через втрату ключових текстурних ознак. Додатково, хоча ансамблева модель демонструє високу загальну точність, хибнопозитивні та хибнонегативні випадки залишаються проблемою, особливо у відео, що використовують найновіші методи генерації діпфейків із покращеною реалістичністю. Крім того, обчислювальна складність ансамблевого підходу може бути проблемою при інтеграції в реальні додатки, де потрібна обробка відео в режимі реального часу.

Можливості для подальших досліджень. Подальші дослідження можуть зосередитися на покращенні стійкості моделі до нових генеративних методів, таких як відео, створені за допомогою StyleGAN3 або Diffusion Models. Перспективним напрямком є інтеграція додаткових ознак, зокрема аудіоаналізу, що дозволить виявляти невідповідності між відеорядом і голосом. Також важливим напрямком розвитку є оптимізація обчислювальних витрат, що може бути досягнуто шляхом використання знань переносного навчання або компресії моделей без суттєвої втрати точності. Додатково, розширене тестування на незалежних наборах даних із реальних джерел допоможе оцінити узагальнюваність моделі для

використання в реальних сценаріях, таких як фактчекінг та цифрова судова експертиза.

10.5 Висновки з проведеного дослідження

Запропонований нейромережевий ансамблевий метод для класифікації діпфейків із використанням вибору золотих кадрів продемонстрував високу ефективність у виявленні підроблених відео. Поєднання базових моделей ResNet50, EfficientNetB0 та Xception у складі ансамблю з метамоделлю XGBoost забезпечило найвищу точність серед протестованих методів, досягнувши 91% на валідаційній вибірці, що перевершує результати Random Forest (82.5%) та Gradient Boosting (85%). Аналіз важливості ознак показав, що найбільший внесок у класифікацію зробили середні значення прогнозів Xception (5.1123) та ResNet (4.8745), тоді як інші ознаки мали нижчу значущість. Додатковий аналіз активацій моделей за допомогою Grad-CAM підтвердив, що різні архітектури нейромереж орієнтуються на різні аспекти обличчя, що робить їхню комбінацію ефективною для детекції аномалій, характерних для діпфейків. Метод вибору золотих кадрів забезпечив значне скорочення обчислювальних витрат, що є критичним фактором для застосування в реальних умовах, зберігаючи при цьому високий рівень точності та узагальнення моделі.

Практичне застосування розробленої системи можливе в декількох ключових сферах. Передусім, вона може бути інтегрована у платформи фактчекінгу та медіа-моніторингу для перевірки автентичності відеоматеріалів, що використовуються в новинах та соціальних мережах. Це дозволить журналістам та незалежним дослідникам оперативно оцінювати достовірність контенту, запобігаючи поширенню дезінформації. Крім того, система може знайти застосування в цифровій судовій

експертизі, допомагаючи виявляти підроблені відеодокази у кримінальних справах. У майбутньому можлива інтеграція цього підходу у системи автоматичного моніторингу відеопотоків, що дозволить виявляти маніпуляції в реальному часі та підвищити рівень інформаційної безпеки.

Попри отримані результати, дослідження виявило певні виклики, зокрема чутливість моделі до якості вхідних відео та складність обробки новітніх генеративних методів. Подальші дослідження мають бути спрямовані на покращення узагальнюваності моделі, інтеграцію аудіоаналізу та оптимізацію обчислювальних ресурсів для забезпечення ефективної роботи в реальному часі. Результати цієї роботи можуть бути корисними для розробки інструментів цифрової судової експертизи, медіа-фактчекінгу та забезпечення інформаційної безпеки в умовах зростаючих загроз, пов’язаних із діпфейк-технологіями.

Список використаних джерел:

1. Teneo. (2024). Deepfakes in 2024 are Suddenly Deeply Real: An Executive Briefing on the Threat and Trends. Retrieved from https://www.teneo.com/insights/articles/deepfakes-in-2024-are-suddenly-deeply-real-an-executive-briefing-on-the-threat-and-trends/
2. Centre for Strategic & International Studies (CSIS). (2024). The Future of Hybrid Warfare. Retrieved from https://www.csis.org/analysis/future-hybrid-warfare
3. Brookings. (2024). Deepfakes and international conflict. Retrieved from https://www.brookings.edu/wp-content/uploads/2023/01/FP_20230105_deepfakes_international_conflict.pdf
4. Geneva Centre for Security Policy (GCSP). (2024). The War in Ukraine: Reality Check for Emerging Technologies and the Future of Warfare.

Retrieved from https://www.gcsp.ch/publications/war-ukraine-reality-check-emerging-technologies-and-future-warfare

5. RAND Corporation. (2024). Ukraine's Lessons for the Future of Hybrid Warfare. Retrieved from https://www.rand.org/pubs/commentary/2022/11/ukraines-lessons-for-the-future-of-hybrid-warfare.html

6. Tran, V.-N., Lee, S.-H., Le, H.-S., & Kwon, K.-R. (2021). High Performance DeepFake Video Detection on CNN-Based with Attention Target-Specific Regions and Manual Distillation Extraction. Applied Sciences, 11(16), 7678. https://doi.org/10.3390/app11167678

7. Zhang, J., Cheng, K., Sovernigo, G., & Lin, X. (2022). A Heterogeneous Feature Ensemble Learning based Deepfake Detection Method. In ICC 2022 - IEEE International Conference on Communications (pp. 2084-2089). https://doi.org/10.1109/ICC45855.2022.9838630

8. Nadimpalli, A. V., & Rattani, A. (2023). Facial Forgery-Based Deepfake Detection Using Fine-Grained Features. In 2023 International Conference on Machine Learning and Applications (pp. 2174-2181). https://doi.org/10.1109/ICMLA58977.2023.00328

9. Guan, L., Liu, F., Zhang, R., Liu, J., & Tang, Y. (2023). MCW: A Generalizable Deepfake Detection Method for Few-Shot Learning. Sensors, 23, 8763. https://doi.org/10.3390/s23218763

10. Khan, R., Sohail, M., Usman, I., Sandhu, M., Raza, M., Yaqub, M. A., & Liotta, A. (2024). Comparative study of deep learning techniques for DeepFake video detection. ICT Express. https://doi.org/10.1016/j.icte.2024.09.018

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

comprehensive survey and analysis. Computer Science Review, 54, 100677. https://doi.org/10.1016/j.cosrev.2024.100677

12. Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications, 252(Part B), 124260. https://doi.org/10.1016/j.eswa.2024.124260

13. Casu, M., Guarnera, L., Caponnetto, P., & Battiato, S. (2024). GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. Forensic Science International: Digital Investigation, 50, 301795. https://doi.org/10.1016/j.fsidi.2024.301795

14. Firc, A., Malinka, K., & Hanáček, P. (2023). Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. Heliyon, 9(4), e15090. https://doi.org/10.1016/j.heliyon.2023.e15090

15. Lee, E.-G., Lee, I., & Yoo, S.-B. (2023). ClueCatcher: Catching Domain-Wise Independent Clues for Deepfake Detection. Mathematics, 11, 3952. https://doi.org/10.3390/math11183952

16. Naitali, A., Ridouani, M., Salahdine, F., & Kaabouch, N. (2023). Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. Computers, 12, 216. https://doi.org/10.3390/computers12100216

17. Dincer, S., Ulutas, G., Ustubioglu, B., Tahaoglu, G., & Sklavos, N. (2024). Golden ratio based deep fake video detection system with fusion of capsule networks. Computers and Electrical Engineering, 117, 109234. https://doi.org/10.1016/j.compeleceng.2024.109234

18. Alarfaj, F.K., & Khan, J.A. (2023). Deep Dive into Fake News Detection: Feature-Centric Classification with Ensemble and Deep Learning Methods. Algorithms, 16, 507. https://doi.org/10.3390/a16110507

19. Koonce, B., & Koonce, B. (2021). ResNet 50. Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, 63-72.

20. Kansal, K., Chandra, T. B., & Singh, A. (2024). ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning" Session id: ICMLDsE. 004". Procedia Computer Science, 235, 70-80.

21. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

22. Xia, X., Xu, C., & Nan, B. (2017, June). Inception-v3 for flower classification. In 2017 2nd international conference on image, vision and computing (ICIVC) (pp. 783-787). IEEE.

23. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

24. Peng, C., Liu, Y., Yuan, X., & Chen, Q. (2022). Research of image recognition method based on enhanced inception-ResNet-V2. Multimedia Tools and Applications, 81(24), 34345-34365.

25. Lipianina-Honcharenko, K., Telka, M., & Melnyk, N. (2024). Comparison of ResNet, EfficientNet, and Xception architectures for deepfake detection. The 1st International Workshop on Advanced Applied Information Technologies CEUR-WS. Khmelnytskyi, Ukraine, Zilina, Slovakia, December 5, 2024. p. 26-34. https://ceur-ws.org/Vol-3899/paper3.pdf

26. Kaggle. (2020). Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/data

Наукове видання

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ В УКРАЇНСЬКИХ ТЕКСТОВИХ ДАНИХ

Колективна монографія

Голова редакційної колегії Ліп'яніна-Гончаренко Х. В

Підписано до друку 23.05.2025 р.
Формат 60x84/16. Папір офсетний.
Друк офсетний. Зам. № 25-259
Умов.-друк. арк. 11,5. Обл.-вид. арк. 12,3.
Тираж 100 прим.

Видавець Західноукраїнський національний університет
вул. Львівська, 11, м. Тернопіль 46009

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців ДК № 7284 від 18.03.2021 р.

Видавничо-поліграфічний центр «Університетська думка»
вул. Бережанська, 2, м. Тернопіль 46009
тел. (0352) 47-58-72
E-mail: edition@wunu.edu.ua