МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЗАХИСТУ ІНФОРМАЦІЇ
КАФЕДРА УПРАВЛІННЯ ІНФОРМАЦІЙНОЮ ТА КІБЕРНЕТИЧНОЮ БЕЗПЕКОЮ**

КВАЛІФІКАЦІЙНА РОБОТА

на тему: “ РОЗРОБКА СИСТЕМИ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ФЕЙКІВ У ТЕКСТОВОМУ КОНТЕНТІ”

на здобуття освітнього ступеня бакалавра
зі спеціальності 125 Кібербезпека
освітньої програми Управління інформаційною та кібернетичною безпекою

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Олексій КЛІМЧЕНКО
(підпис) *Ім’я, ПРІЗВИЩЕ здобувача*

Виконав: здобувач вищої освіти гр. УБД-41

Олексій КЛІМЧЕНКО
Ім’я, ПРІЗВИЩЕ

Керівник: Віталій ТИЩЕНКО
Ім’я, ПРІЗВИЩЕ

Рецензент:
К.т.н., доцент Ім’я, ПРІЗВИЩЕ

Київ 2024

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут захисту інформації

Кафедра управління інформаційною та кібернетичною безпекою

Ступінь вищої освіти бакалавр

Спеціальність 125 Кібербезпека

Освітня програма Управління інформаційною та кібернетичною безпекою

ЗАТВЕРДЖУЮ
Завідувач кафедри УІКБ
_____ Світлана ЛЕГОМІНОВА
“____”_____2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Клімченку Олексію Романовичу
(прізвище, ім’я, по батькові здобувача)

1. Тема кваліфікаційної роботи “Розробка системи машинного навчання для виявлення фейків у текстовому контенті”,
керівник кваліфікаційної роботи ТИЩЕНКО Віталій.
(ІПРІЗВИЩЕ, Ім’я,, науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологій "Про закріплення тем випускних кваліфікаційних робіт та призначення наукових керівників на 2023-2024 н.р. за студентами першого (бакалаврського) рівня вищої освіти". № 36 від 27.02.24

2. Строк подання кваліфікаційної роботи “20” травня 2024р.

3. Вихідні дані до кваліфікаційної роботи: *методи машинного навчання, інструменти аналізу тексту, методи та засоби забезпечення інформаційної безпеки, міжнародні стандарти, наукова та технічна література.*

4. Перелік питань, які мають бути розроблені:
   4.1. Проаналізувати поняття фейків у текстовому контенті.
   4.2. Дослідити методи виявлення фейків у текстовому контенті.
   4.3. Вивчити процес розробки системи машинного навчання.

5. Перелік ілюстративного матеріалу: *презентація PowerPoint*

6. Дата видачі завдання “11“ березня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Етапи кваліфікаційної роботи	Термін виконання етапів роботи	Примітка
1.	Визначення об’єкту, предмету, мети та завдань дослідження.	18.03.2024	
2.	Збір та аналіз літератури.	29.03.2024	
3.	Аналіз поняття фейків у текстовому контенті.	08.04.2024	
4.	Дослідження методів виявлення фейків у текстовому контенті.	22.04.2024	
5.	Вивчення процесу розробки системи машинного навчання.	08.05.2024	
6.	Формулювання висновків за результатами проведеного дослідження.	20.05.2024	
7.	Оформлення роботи.	22.05.2024	
8.	Оформлення презентації.	03.06.2024	
9.	Отримання рецензії на роботу.	03.06.2024	
10.	Захист в ДЕК.	__.06.2024	

Здобувач вищої освіти _____ Олексій КЛІМЧЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Керівник
кваліфікаційної роботи _____ Віталій ТИЩЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЗАХИСТУ ІНФОРМАЦІЇ

ПОДАННЯ ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня бакалавра

Направляється здобувач Клімченко О.Р. до захисту кваліфікаційної роботи
(прізвище та ініціали)
за спеціальністю 125 Кібербезпека
(код, найменування спеціальності)
освітньої програми Управління інформаційною та кібернетичною безпекою
(назва)
на тему: “Розробка системи машинного навчання для виявлення фейків у текстовому контенті”

Кваліфікаційна робота і рецензія додаються.

Директор ННІЗІ _____ Віталій САВЧЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувач КЛІМЧЕНКО Олексій у кваліфікаційній роботі проаналізував особливості розробки системи машинного навчання для виявлення фейків у текстовому контенті, дослідив основні методи та технології обробки природної мови, вивчив алгоритми машинного навчання для класифікації тексту, розробив практичні рекомендації за темою дослідження.

КЛІМЧЕНКО Олексій показав розуміння проблеми дослідження та бачення основних теоретичних і практичних напрямів її вирішення, довів володіння методами наукового дослідження, проявив себе як організований, відповідальний виконавець. Результати дослідження апробовані на двох конференціях.

Все це дозволяє оцінити кваліфікаційну роботу здобувача КЛІМЧЕНКА Олексія на оцінку “_____” та присвоїти йому кваліфікацію бакалавра з кібербезпеки за освітньою програмою Управління інформаційною та кібернетичною безпекою.

Керівник кваліфікаційної роботи _____ Віталій ТИЩЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

“____“_____2024 року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувач Клімченко О.Р. допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедри
управління інформаційною
та кібернетичною безпекою _____ Світлана ЛЕГОМІНОВА
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну бакалаврську роботу

здобувача вищої освіти КЛІМЧЕНКА Олексія
на тему “Розробка системи машинного навчання для виявлення фейків у текстовому контенті”

Актуальність. У сучасному світі, де зростає залежність суспільства від інформаційних технологій та кіберпростору, тема розробки системи машинного навчання для виявлення фейків у текстовому контенті є надзвичайно актуальною. Поширення фейкових новин та дезінформації може мати серйозні наслідки для громадської думки, політичної стабільності та економічного благополуччя. Тому розробка ефективних методів для автоматичного виявлення фейкових новин є критично важливою для забезпечення достовірності інформації та захисту суспільства від маніпуляцій.

Позитивні сторони.

1. В роботі використано сучасні методи машинного навчання, що дозволяють ефективно виявляти фейкові новини. Це свідчить про високий рівень теоретичної підготовки здобувача.
2. Розроблена система має значний потенціал для практичного застосування, зокрема, у засобах масової інформації та соціальних мережах для автоматичного фільтрування контенту.
3. Здобувач провів всебічний аналіз існуючих методів і підходів до виявлення фейків, що дозволило вибрати оптимальне рішення для розробки власної системи.
4. Прототип системи показав високу ефективність на тестових даних, що підтверджує правильність обраних методів та підходів.
5. Робота відзначається високою якістю оформлення, логічною структурою та чіткістю викладу матеріалу.

Недоліки.

1. Тестування системи проводилося на відносно невеликому наборі даних, що може не повністю відображати її ефективність в реальних умовах.
2. Існує потенціал для подальшого вдосконалення алгоритмів з метою підвищення

Висновок: Кваліфікаційна робота виконана на належному науково-методичному рівні і заслуговує оцінки “_____”, а здобувач КЛІМЧЕНКО Олексій заслуговує присвоєння кваліфікації бакалавра з кібербезпеки за освітньою програмою Управління інформаційною та кібернетичною безпекою.

Рецензент: _____ _____
підпис Ім’я, ПРІЗВИЩЕ

РЕФЕРАТ

Кваліфікаційна робота присвячена розробці системи машинного навчання для виявлення фейків у текстовому контенті. Робота складається зі вступу, трьох розділів, що містять 7 рисунків, висновків і списку використаних джерел із 47 найменувань. Загальний обсяг роботи становить 69 аркушів, з яких 5 аркуші займають перелік умовних скорочень та список використаних джерел.

Метою роботи є створення ефективної системи машинного навчання для автоматичного виявлення фейкових новин у текстових даних.

Об'єктом дослідження є текстовий контент, який аналізується на наявність фейків.

Предмет дослідження **–** методи та технології машинного навчання, застосовані для виявлення фейків у текстовому контенті.

Методи дослідження. Для вирішення означеного наукового завдання в роботі використані методи аналізу та синтезу, машинного навчання, обробки природної мови (NLP), порівняння, класифікації, та експертної оцінки.

Як результат, у роботі проаналізовано сучасні методи машинного навчання, досліджено технології обробки природної мови для виявлення фейкових новин, розроблено та протестовано модель машинного навчання, запропоновано практичні рекомендації щодо її застосування.

Галузь застосування. Розроблені підходи можуть бути використані для автоматичного виявлення фейкових новин у текстових даних в медіа, соціальних мережах, та інших інформаційних ресурсах.

Ключові слова: МАШИННЕ НАВЧАННЯ, ВИЯВЛЕННЯ ФЕЙКІВ, ТЕКСТОВИЙ КОНТЕНТ, ОБРОБКА ПРИРОДНОЇ МОВИ, НЕЙРОННІ МЕРЕЖІ, КЛАСИФІКАЦІЯ ТЕКСТУ.

ABSTRACT

The qualification work is devoted to the development of a machine learning system for detecting fakes in textual content. The work consists of an introduction, three chapters containing 7 figures, conclusions and the list of references containing 47 items. The total volume of the work is 69 pages, of which 5 pages are occupied by the list of abbreviations and the list of references.

The purpose of the study is to create an effective machine learning system for automatically detecting fake news in text data.

The object the study is textual content that is analyzed for fake news.

The subject of the study is machine learning methods and technologies used to detect fake news in text content.

Research methods. In order to solve the mentioned higher scientific task, the methods of analysis and synthesis, machine learning, natural language processing (NLP), comparison, classification, and expert evaluation.

As a result, the work analyzed modern machine learning methods, explores natural language processing technologies for detecting fake news, develops and tests a machine learning model, and offers practical recommendations for its application.

Field of application. The developed approaches can be used to automatically detect fake news in text data in the media, social networks, and other information resources.

Keywords: MACHINE LEARNING, FAKE NEWS DETECTION, TEXTUAL CONTENT, NATURAL LANGUAGE PROCESSING, NEURAL NETWORKS, TEXT CLASSIFICATION.

ЗМІСТ

ВСТУП

Актуальність теми. У сучасному світі, де зростає залежність суспільства від інформаційних технологій та кіберпростору, тема розробки системи машинного навчання для виявлення фейків у текстовому контенті є надзвичайно актуальною. Поширення фейкових новин та дезінформації може мати серйозні наслідки для громадської думки, політичної стабільності та економічного благополуччя. Тому розробка ефективних методів для автоматичного виявлення фейкових новин є критично важливою для забезпечення достовірності інформації та захисту суспільства від маніпуляцій.

Мета роботи полягає у розробці системи машинного навчання для автоматичного виявлення фейків у текстовому контенті.

Об'єкт дослідження – текстовий контент, який аналізується на наявність фейків.

Предмет дослідження методи та технології машинного навчання, застосовані для виявлення фейків у текстовому контенті.

Для досягнення цієї мети в роботі необхідно виконати наступні **завдання**:

1. Проаналізувати поняття фейків у текстовому контенті.
2. Дослідити методи виявлення фейків у текстовому контенті.
3. Вивчити процес розробки системи машинного навчання.

Методи дослідження. ля вирішення означеного наукового завдання в роботі використані методи аналізу та синтезу, машинного навчання, обробки природної мови (NLP), порівняння, класифікації, та експертної оцінки.

Як результат, у роботі проаналізовано сучасні методи машинного навчання, досліджено технології обробки природної мови для виявлення фейкових новин, розроблено та протестовано модель машинного навчання, запропоновано практичні рекомендації щодо її застосування.

Практичне значення одержаних результатів. Застосування напрацювань дозволить ефективно виявляти фейкові новини в текстовому контенті, що сприятиме покращенню інформаційної безпеки та достовірності

новин. Результати дослідження можуть бути використані для оптимізації систем автоматичного аналізу тексту, спираючись на оцінку існуючих методів та рекомендації щодо їх покращення. Це дозволить підвищити якість інформації, що розповсюджується в медіа та соціальних мережах, та захистити суспільство від негативного впливу фейкових новин.

Апробація результатів кваліфікаційної роботи відбулася на Всеукраїнській науково-практичній конференції “Стратегії кіберстійкості: управління ризиками та безперервність бізнесу” 28 лютого 2024 року.

РОЗДІЛ 1 ПОНЯТТЯ ФЕЙКІВ У ТЕКСТОВОМУ КОНТЕНТІ

1.1 Поняття фейків в цифровому середовищі

У часи інформаційної ери мережа Інтернет та соціальні мережі є найбільш сприятливим для створення та поширення всеможливого контенту. З огляду на це, кожна людина може вільно опублікувати будь-яку інформацію, яка свідомо або випадково виявиться неправдивою, активне поширення фейків є очевидним негативним наслідком. Дезінформація не є новим поняттям в сучасному світі – з давніх часів люди використовували її у своїх цілях для маніпулювання людьми. З розвитком у 20-му столітті технологій і появою нових засобів масової інформації, таких як радіо та телебачення, ризик натрапити на фейк підвищувався з кожним роком.

Наразі в цифровому середовищі вся інформація поширюється з неймовірною швидкістю та всебічно впливає на громадську думку. За останні 10 років фейки стали часто фігурувати на рівні з правдивими новинами, що призводить до погіршення критичного аналізу інформації, що споживає людство. У 2017 році «fake news» стало словом року за Collins English Dictionary [1]. Також Google Trends показує різкий зріст інтересу до "фейкових новин" у листопаді 2016 року та на початку 2020 року (Рис. 1). Як можемо спостерігати, тенденція пошуку терміну фейкових новин не спадає.

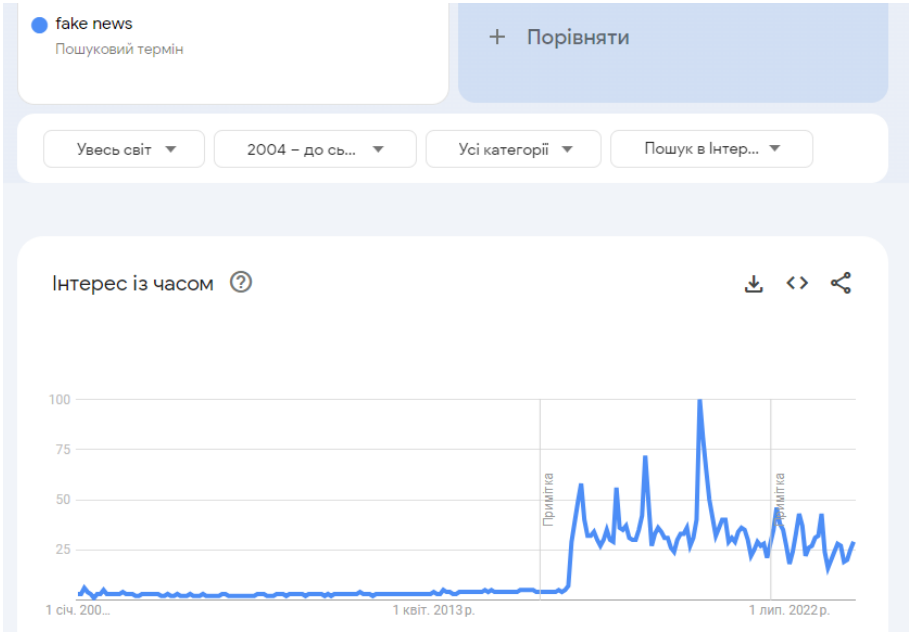

Рис. 1.1. Дані Google Trends про пошук «fake news» за останні 20 років

У літературі існує безліч термінів і понять, які використовуються для позначення неправдивої, недостовірної або напівправдивої інформації: "фейкові новини", "неправдиві новини", "цифрова дезінформація", "дезінформація", "чутки" тощо. Для кращого розуміння різниці, розглянемо найпоширеніші терміни, що використовуються найчастіше як синоніми: «misinformation», «disinformation» та «fake news».

Найкоректнішим перекладом на українську для терміну «misinformation» буде «спотворена інформація». Цим терміном визнвчається неправдива інформація, яка поширюється незалежно від наміру ввести в оману. Дезінформація - це свідоме поширення неправдивої інформації з метою обману, введення в оману або впливу на громадську думку. Це може бути частиною систематичної кампанії або маніпуляції, спрямованої на досягнення певних цілей, таких як політична або соціальна маніпуляція, дестабілізація суспільства тощо. Термін фейкові новини використовується для опису неправдивої інформації, яка прикрашена або вигадана повністю з метою привернути увагу, збільшити аудиторію або вплинути на громадську думку.

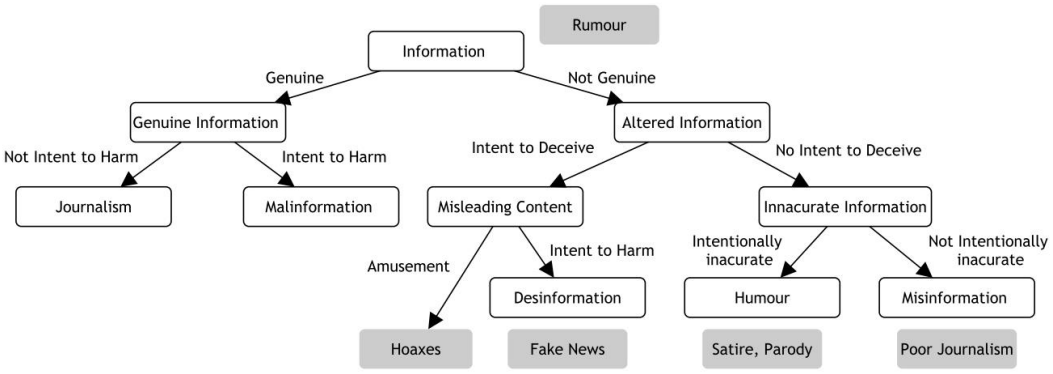

Рис. 1.2. Класифікація фейкової інформації [2]

Новина – це повідомлення про певну реальну і нещодавню подію, що становить загальний інтерес, яка так чи інакше впливає на суспільство і передає щось нове його аудиторії. Новини – це продукт, який є результатом роботи журналістів, які, згідно з основними принципами журналістики, дотримуються принципів достовірності та об'єктивності у висвітленні фактів. Тому поняття "фейкові новини", в строгому сенсі слова, є суперечливим.

Згідно з визначенням David Buckingham, "фейкові новини - це новини, які сфабриковані і навмисно призначені для введення в оману, зазвичай з'являються на сайтах, які маскуються під справжні новинні сайти" [3]. Цей підхід до розповсюдження неправдивої інформації є надзвичайно небезпечним, оскільки фальсифіковані новини можуть виглядати цілком достовірними для пересічного читача.

Інше визначення пропонує Don Fallis, який зазначає, що "фейкові новини - це підроблені новини. Матеріал є фейковою новиною тоді і тільки тоді, коли він не є справжньою новиною, але подається як справжня з наміром і прагненням ввести в оману" [4]. Це означає, що фейкові новини цілеспрямовано створюються для маніпуляції громадською думкою, що може мати серйозні наслідки для суспільства.

Jessica Pepp визначає фейкові новини як "широкорозповсюдження матеріалів, які розглядаються особами, що їх поширюють, як такі, що будуть отримані за допомогою стандартних журналістських практик, але насправді

такими практиками не були створені" [5]. Це визначення підкреслює, що фейкові новини часто імітують стиль і форму традиційної журналістики, що робить їх особливо складними для виявлення.

Edson C. Tandoc Jr. зазначає, що "фейкові новини прикидаються виглядом і сенсом справжніх новин, автори яких мають на меті ввести в оману, мають низький рівень достовірності і високий рівень прямого наміру ввести в оману" [6]. Таким чином, фейкові новини характеризуються не лише неправдивістю, але й свідомим наміром дезінформувати аудиторію.

Як можемо спостерігати, більшість авторів сходяться на думці, що фейкові новини – це публікації, які імітують формат новин або репортажів, з неправдивими твердженнями або інформацією, створеною з метою ввести в оману або маніпулювання читачем.

Окрім базового розподілу на дезінформацію, спотворену інформацію та фейкові новини, що ми розглянули раніше, можна класифікувати фейки за іншими ознаками, що включають, але не обмежують, методи створення, поширення та мету розповсюдження.

Методи створення фейків можуть варіюватися від використання програмного забезпечення для фотошопу та відеомонтажу до застосування штучного інтелекту для генерації фальшивих матеріалів, що робить їх надзвичайно важкими для виявлення. Маніпуляції з реальними зображеннями або відео для створення неправдивих новин стають все більш поширеними завдяки новим технологіям. Привласнення фотографій є ще одним видом таких фейкових новин: фотографію, вирвану з контексту, навмисно чи ненавмисно приписують до історії, яка не має до неї відношення, чим вводять в оману людей. [7]

Також важливо враховувати методи поширення, які можуть включати в себе використання соціальних мереж, месенджерів або електронної пошти. Громадянська журналістика – блоги, авторські колонки та інший контент, створений громадянами для поширення інформації звичайними людьми, яка є емоційно забарвленою і не відповідає журналістським нормам, тому її можна

віднести до фейкових новин. Реклама, замаскована під справжні новини або інформацію, також може розглядатися як фейк. Також існує поняття клікбейту – це навмисне використання оманливих заголовків, щоб заохотити відвідувачів натиснути на певну веб-сторінку або посилання. Багато з цих кліків переводять читача на комерційний сайт, а не на новинний, так як бажання дізнатися про сенсаційні та дискусійні теми приваблює людей і вводить їх в оману з корисливою метою. Заробіток на рекламі через глядацьку аудиторію або фішинг є одними з основних цілей цього типу фейкових новин.

Мета розповсюдження фейків може бути різноманітною, від політичної маніпуляції до залучення уваги або навіть виклику паніки серед населення. Пропаганда – несправедливо упереджена та оманлива інформація, що поширюється в цільових спільнотах відповідно до заздалегідь визначеної стратегії для просування певної точки зору або політичного порядку денного, та яка спрямована на отримання політичної або фінансової вигоди. [8]

1.2 Машинне навчання

З часів своєї еволюції люди використовували різні види інструментів, щоб виконувати різні завдання у більш простий спосіб. Творчий потенціал людського мозку призвів до винайдення різноманітних машин. Галузь машинного навчання розвинулася з розширеної галузі штучного інтелекту, яка має на меті імітувати інтелектуальні здібності людини за допомогою машин. Машинне навчання (ML) використовується для того, щоб навчити машини більш ефективно обробляти дані. Машинне навчання зазвичай відноситься до різновидів пристроїв, які виконують ролі, пов'язані зі штучним інтелектом (ШІ), такі як розпізнавання, планування, управління роботами, прогнозування тощо. [9]

Машинне навчання (МН) – це класифікація обчислень, яка дозволяє програмним забезпеченням ставати все більш точними в прогнозуванні результатів без потреби в чіткому налаштуванні. Основна ідея МН полягає у

створенні обчислень, які можуть отримувати вхідну інформацію та використовувати фактичні дослідження для прогнозування результатів, оновлюючи їх у міру надходження нової інформації [10].

Алгоритми машинного навчання поділяються на п'ять широких категорій: контрольоване навчання, неконтрольоване навчання, напівконтрольоване навчання, самоконтрольоване навчання та навчання з підсиленням. До контрольованого навчання відносяться регресійні алгоритми, алгоритми класифікації, наївні класифікатори Байєса, нейронні мережі та алгоритми випадкових дерев. неконтрольоване навчання в той час ділиться на кластеризацію K-середніх, ієрархічну кластеризацію, імовірнісну кластеризацію. Розглянемо кожен з алгоритмів більш детально. [11]

Лінійна регресія - це метод, який використовується в статистичному моделюванні для визначення зв'язку між кількома змінними. Регресія досліджує, як змінюється залежна змінна при зміні однієї незалежної змінної, тоді як інші незалежні змінні залишаються незмінними. Лінійна регресія буває двох типів: Проста лінійна регресія та множинна лінійна регресія Логістична регресія дуже схожа на лінійну регресію. Лінійна регресія використовується для вирішення питань регресії, тоді як логістична регресія використовується для вирішення завдань класифікації. Логістична регресія - це керована техніка машинного навчання для прогнозування ймовірності певного класу або події. Вона використовується, коли дані розділені лінійно, а результат є бінарним або категоріальним. Тобто, логістична регресія зазвичай застосовується до питань бінарної класифікації. Оцінювання вихідної змінної, яка є дискретною у двох класах, називається бінарною класифікацією. [12]

Алгоритми класифікації – прогнозують категоричні вихідні змінні (наприклад, "сміття" або "не сміття") шляхом маркування фрагментів вхідних даних. Алгоритми класифікації включають логістичну регресію, метод k-найближчих сусідів та машини опорних векторів. Наївні класифікатори Байєса (NB) - це метод класифікації, заснований на теоремі Байєса [13]. Ця теорема може описати ймовірність події на основі попередніх знань про обставини,

пов'язані з цією подією. Цей класифікатор припускає, що певна ознака в класі не пов'язана безпосередньо з будь-якою іншою ознакою, хоча ознаки для цього класу можуть мати взаємозалежність між собою. Нейронні мережі імітують роботу людського мозку з величезною кількістю пов'язаних між собою вузлів обробки, які можуть полегшити такі процеси, як переклад природної мови, розпізнавання зображень, розпізнавання мови і створення зображень.

Метод випадкових дерев використовує безліч дерев рішень для створення "лісу", який є більш точним і надійним, ніж його окремі компоненти. Кожне дерево рішень у методі випадкових дерев приймає рішення на основі заданих даних. Коли ці дерева працюють разом, вони утворюють "ліс", який може впоратися зі складними проблемами набагато краще, ніж будь-яке окреме дерево. Алгоритм випадкових дерев допомагає уникнути таких поширених проблем, як надмірне налаштування моделі.

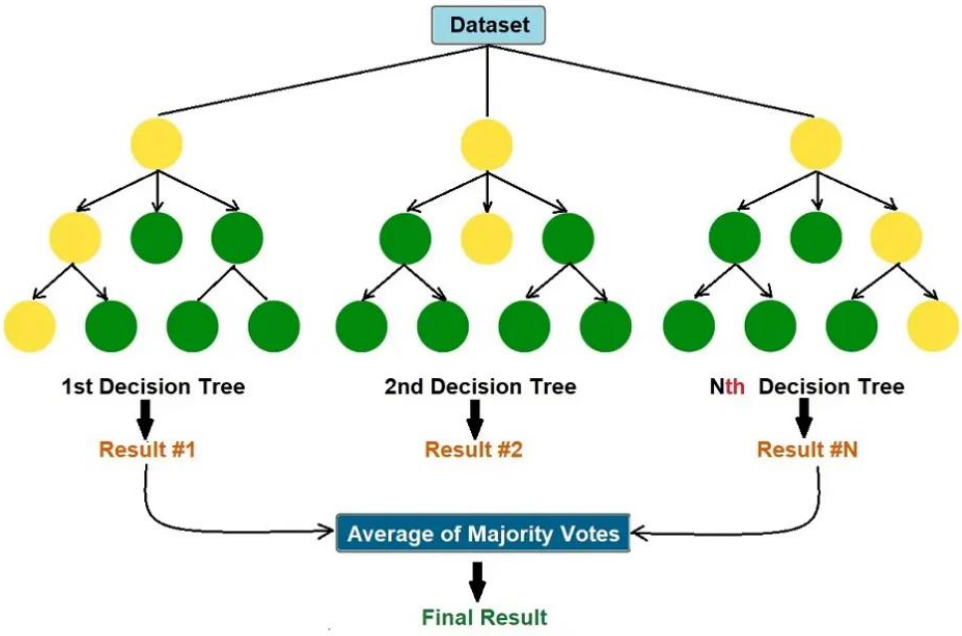

Рис. 1.3. Алгоритм випадкових дерев

Кластеризація K-середніх - розподіляє точки даних на K груп, де точки даних, найближчі до заданого центроїда, об'єднуються в одну категорію, а K представляє кластери на основі їхнього розміру та рівня деталізації. Кластеризація за методом K-середніх зазвичай використовується для

сегментації ринку, кластеризації документів, сегментації зображень і стиснення зображень.

Ієрархічна кластеризація – описує набір методів кластеризації, включаючи агломеративну кластеризацію, коли точки даних спочатку ізолюють у групи, а потім ітеративно об'єднують на основі схожості, поки не залишиться один кластер, і дивізіональну кластеризацію, коли один кластер даних розділяється на основі відмінностей між точками даних.

Імовірнісна кластеризація – допомагає вирішувати проблеми оцінки щільності або "м'якої" кластеризації, групуючи точки даних на основі ймовірності того, що вони належать до певного розподілу.

Самонавчання отримує керуючі сигнали з безпосередньо самих даних, часто використовуючи базову структуру даних. Загальна методика самонавчання полягає в тому, щоб передбачити будь-яку неспостережувану або приховану частину (або властивість) вхідних даних за будь-якою спостережуваною або прихованою частиною вхідних даних. [14]

Навчання з підсиленням, яке також називають навчанням з використанням зворотного зв'язку з людиною (RLHF), – це тип динамічного програмування, який тренує алгоритми за допомогою системи заохочень і покарань. Щоб застосувати навчання з підсиленням, агент виконує дії в певному середовищі для досягнення заздалегідь визначеної мети. Агент отримує винагороду або покарання за свої дії на основі встановленої метрики (зазвичай у балах), заохочуючи агента використовувати хороші практики і відмовлятися від поганих. З повторенням агент навчається найкращим стратегіям.

Напівкероване навчання - це один з видів машинного навчання (ML). Він знаходиться на проміжній стадії між контрольованим і неконтрольованим навчанням, тобто набір даних частково маркується. Основною метою SSL є подолання недоліків як контрольованого, так і неконтрольованого навчання. Контрольоване навчання вимагає величезної кількості навчальних даних для класифікації тестових даних, що є економічно ефективним і трудомістким процесом. З іншого боку, неконтрольоване навчання не вимагає жодних

маркованих даних, а кластеризує дані на основі схожості точок даних, використовуючи або кластеризацію, або метод максимальної правдоподібності. Основний недолік цього підходу полягає в тому, що він не може точно кластеризувати невідомі дані. [15]

Існує кілька процесів, які можна реалізувати за допомогою машинного навчання, в яких відбувається включення кластерів, таких як ієрархічні кластери, неієрархічні кластери і т.д., що дозволяє залучити машинне навчання поряд зі штучним інтелектом. Завдяки появі технологій, штучний інтелект отримав розвиток у кількох аспектах, таких як виявлення помилок, розробка програмного забезпечення, штучні нейронні мережі та інші подібні сфери. Системи виявлення помилок у програмному забезпеченні є добре відомим процесом, коли йдеться про сучасну розробку програмного забезпечення, оскільки зростання кількості помилок і дефектів дозволяє системі вчитися і уникати повторення помилок. [16]

Ще одна причина, чому нейронні мережі змінюють підхід до роботи, полягає в тому, що вони масштабуються. Традиційні програми стають більш масштабними, коли вони стають більшими – їх вже не так легко контролювати, і вони схильні до багів (помилок) або вразливостей у системі безпеки. Нейронні мережі можуть додавати більше рівнів елементів без ускладнення [17]. Цілком можливо, що в майбутньому значна частина всього програмного забезпечення буде замінена нейронними мережами.

1.3. Хто досліджував

Від початку існування людського суспільства і до сьогодні людство неминуче стикається з маніпуляціями та фейками, спрямованими на вплив на громадську думку та спотворення подій. Дослідження фейкових новин у текстовому середовищі охоплює низку дисциплін, зокрема літературознавство, лінгвістику та соціологію.

Вальтер Беньямін – німецький інтелектуал єврейського походження, літературний критик, філософ, соціолог, перекладач, радіоведучий та есеїст. Його праці, що поєднують історичний матеріалізм, німецький ідеалізм та єврейський містицизм, здійснили значний і впливовий внесок у естетичні теорії та західний марксизм, їх часто асоціюють із франкфуртською школою критичної теорії. [18]

Основними темами есе Беньяміна "Витвір мистецтва в епоху механічного відтворення" є взаємозв'язок між твором мистецтва, технологіями його відтворення та політичними змінами. Технологічні зміни, які уможливлюють відтворення численних копій творів мистецтва, кидають виклик традиційним уявленням про твір мистецтва як автономний об'єкт, наділений "аурою" автентичності та незмінної цінності. Беньямін стверджує, що з масовим поширенням зображень "аура", притаманна твору мистецтва, розмивається. На думку Беньяміна, кіно та фотографія вплинули на те, як ми сприймаємо та відчуваємо світ. Це пов'язано з тим, що очевидна відсутність автентичності фільму – той факт, що немає нічого, щоб відрізнити "оригінальну" плівку від її копій – робить аудиторію більш уважною до інших штучних проявів. [19]

Дослідження Габріеля Зорана про вплив технологій на маніпуляції з текстом та фальсифікацію є важливим напрямком цифрової гуманітарної науки. Він досліджує, як технології впливають на створення текстів, зокрема у зв'язку з розвитком штучного інтелекту. Наприклад, обробка природної мови та машинне навчання дозволяють створювати тексти, які імітують різні стилі письма та зміст.

Тексти, створені штучним інтелектом, можуть порушувати поняття авторства та оригінальності, що створює проблеми з автентичністю контенту, створеного людиною. Крім того, Зоран досліджує, як поширення та доступність текстів змінюються в цифрову епоху. Наприклад, на платформах соціальних мереж можуть поширюватися фейкові тексти без належної перевірки.

Ці дослідження стикаються з проблемами, з якими зіштовхуються як науковці, так і читачі. Для науковців поширення фейкових текстів ускладнює

пошук та перевірку матеріалів для досліджень, а для читачів може бути складно розпізнати автентичність текстів, з якими вони зіштовхуються, що призводить до дезінформації та зменшення довіри до цифрового контенту.

В роботах Зорана також розглядаються етичні та правові аспекти. Маніпуляції з текстом ставлять під загрозу авторські права, інтелектуальну власність та цілісність наукової та журналістської роботи. Правові рамки часто не встигають за технологічним прогресом, що створює прогалини у захисті від цифрового підроблення. [20]

Загалом, дослідження Зорана підкреслюють важливість бути пильними, критично мислити та розробляти надійні інструменти для перевірки та підтвердження автентичності текстів у цифровому світі. Вони також показують постійну боротьбу між технологічними інноваціями та збереженням довіри та автентичності в цифрових комунікаціях.

Ролан Барт, ключова фігура в літературній теорії та семіотиці 20-го століття, глибоко вплинув на те, як ми розуміємо тексти, авторство та значення. Його есе "Смерть автора" пропонує фундаментальну концепцію, яка перетинається з сучасними проблемами текстової автентичності та цифрової маніпуляції текстами.

Згідно з Бартом, значення тексту не залежить від намірів його оригінального автора, а натомість конструюється читачами, які з ним взаємодіють. Ця концепція розвіює містику навколо авторської "аури" і віддає інтерпретаційну владу в руки аудиторії, припускаючи, що текст приховує в собі безліч значень, жодне з яких не може вважатися остаточним або авторитетним. Теорія Барта виступає за відхід від зосередження уваги на оригінальному авторі до акцентування на тому, як тексти сприймаються і розуміються читачами, незалежно від їхнього джерела. Така децентралізація авторського задуму особливо актуальна в дискусіях навколо маніпуляцій з цифровими текстами.

Концепція Барта дає читачам можливість брати активну участь у створенні сенсів, а не пасивно сприймати фіксоване повідомлення. У цифровій сфері це розширення можливостей сприяє критичному та допитливому підходу до

текстів, необхідному для розпізнавання автентичності та дезінформації. Контент, створений штучним інтелектом, анонімні публікації та спільні проекти з написання текстів розмивають межі між автором та аудиторією, посилюючи актуальність ідей Барта.

Підхід Барта створює етичні та інтерпретаційні дискусії, особливо в контекстах, де автентичність і походження тексту мають першорядне значення, як, наприклад, у правовому, академічному та журналістському середовищі. [21]

Роботи Барта пропонують цінну основу для ретельного вивчення складнощів текстів, авторства та автентичності в сучасну епоху. Його ідеї змушують нас переосмислити роль автора в умовах, коли технології здатні маніпулювати текстами, і кидають виклик упередженням щодо оригінальності та значущості.

Висновки до розділу 1

Розділ висвітлює ключові аспекти розуміння, класифікації та методів виявлення фейкової інформації. Аналіз цих питань сприяє глибшому розумінню проблеми дезінформації та можливих шляхів її вирішення.

Визначення фейків полягає у розгляді різних підходів до ідентифікації неправдивої інформації. У сучасному науковому дискурсі фейки розглядаються як інформація, що навмисно створена або поширюється з метою введення в оману. Фейковий контент може мати різні форми, включаючи тексти, зображення, відео та аудіо, і може бути використаний для досягнення політичних, економічних або соціальних цілей. Визначення фейків включає аналіз їх характеристик, таких як відсутність надійних джерел, використання маніпулятивних технік та поширення через сумнівні канали.

Машинне навчання є ключовою технологією у боротьбі з фейковими новинами. Алгоритми машинного навчання використовуються для автоматизованого виявлення неправдивої інформації. Ці алгоритми можуть аналізувати великі обсяги текстового контенту, ідентифікувати патерни,

характерні для фейкових новин, та класифікувати інформацію як правдиву або неправдиву. Використання методів обробки природної мови (NLP) дозволяє алгоритмам аналізувати лінгвістичні особливості тексту, що допомагає у виявленні маніпулятивних або брехливих заяв. Удосконалення моделей машинного навчання постійно покращує їхню здатність розрізняти фейки від правдивих новин, знижуючи ризик поширення дезінформації.

Дослідження фейків проводилися багатьма вченими та дослідницькими групами. Серед найвідоміших дослідників у цій сфері можна виділити групи з Массачусетського технологічного інституту (MIT), Оксфордського університету та Стенфордського університету. Їхні роботи зосереджені на розробці алгоритмів для виявлення фейкових новин, аналізі соціальних мереж як платформи для поширення дезінформації та вивченні впливу фейкових новин на суспільну думку. Ці дослідники використовують міждисциплінарний підхід, поєднуючи знання з інформатики, соціології, психології та лінгвістики для всебічного розуміння проблеми.

Тобто розділ окреслює комплексний підхід до дослідження фейкової інформації, включаючи її визначення, технологічні засоби боротьби та науковий вклад різних дослідників у цю сферу.

РОЗДІЛ 2 МЕТОДИ ВИЯВЛЕННЯ ФЕЙКІВ У ТЕКСТОВОМУ КОНТЕНТІ

2.1. Метод обробки природної мови

Обробка природної мови (Natural Language Processing, NLP) - це спеціалізована галузь штучного інтелекту, яка зосереджується на комп'ютерній інтерпретації та аналізі людської мови. Ця сфера включає широкий спектр напрямків, пов'язаних з обробкою тексту та аудіо, використовуючи методи машинного навчання, що базуються на статистичних підходах. Завдяки впровадженню різних методологій, NLP також охоплює прагматичні аспекти комп'ютерної лінгвістики, які стали надзвичайно розвиненими та потужними.

З кожним роком можливості та доступність методів NLP зростають, що сприяє покращенню точності та ефективності комп'ютерного аналізу мови. Обробка природної мови та машинне навчання залишаються одними з найважливіших і найбільш досліджуваних напрямків у сучасній науці. На розвиток NLP значною мірою впливають інші дисципліни, такі як психологія, когнітивна наука та лінгвістика, оскільки вони надають цінні інсайти для створення обчислювальних моделей, здатних вирішувати завдання, пов'язані з людською взаємодією та розумінням мови [22].

Для підтримки та розвитку цієї галузі було створено кілька програмних пакетів, спрямованих на мовне моделювання та інтерпретацію мови. Ці інструменти допомагають перетворювати комп'ютерну мову на таку, яку легко зрозуміти людині, роблячи взаємодію з машинами більш природною та ефективною.

У рамках NLP використовуються три широкі концепції, які мають фундаментальне значення для цієї галузі:

1. *Суб'єктивність*:

Суб'єктивна репрезентація світу є основною концепцією, що відображає індивідуальний досвід людини. Цей досвід формується завдяки п'яти основним

сенсам: дотику, нюху, смаку, зору та слуху. Кожна з цих сенсорних модальностей відіграє важливу роль у створенні суб'єктивної свідомості, яка є невід'ємною частиною людського сприйняття. У контексті обробки природної мови, цей суб'єктивний досвід можна порівняти з мовною свідомістю, де людина здатна інтерпретувати та взаємодіяти з мовою на основі своїх сенсорних відчуттів. Виходячи з цього, NLP іноді розглядається як дослідження структури суб'єктивного досвіду.

2. *Свідомість*:

Свідомість людського розуму складається з двох ключових компонентів: свідомої та несвідомої частин. Суб'єктивна репрезентація, яка відбувається у свідомості, називається свідомістю. Водночас репрезентації, що відбуваються поза межами свідомості, належать до несвідомого розуму. У NLP свідомість відіграє центральну роль, оскільки багато процесів інтерпретації мови залежать від свідомих та несвідомих когнітивних механізмів [23].

3. *Навчання*:

Когнітивне навчання людини починається з моменту її народження і активно залучає всі п'ять органів чуття. Люди вчаться взаємодіяти зі світом завдяки сенсорним сприйняттям та досвіду, який вони накопичують протягом життя. У контексті NLP, навчання базується на принципах моделювання, де комп'ютеризовані моделі створюються на основі людського досвіду. Ці моделі використовують сенсорні та лінгвістичні репрезентації, що кодифікуються для ефективної обробки мови. Вони спрямовані на створення систем, здатних точно інтерпретувати та реагувати на природну мову людини.

Обробка природної мови інтегрує суб'єктивний досвід, свідомість та навчання для створення моделей, які можуть взаємодіяти з людською мовою на високому рівні. Ці концепції допомагають розширити можливості NLP, роблячи його інструментом, що здатен забезпечити більш природну та ефективну комунікацію між людиною і машиною.

NLP включає кілька ключових етапів, кожен з яких відіграє важливу роль у забезпеченні точного та ефективного аналізу мови. Ці етапи охоплюють

морфологічний, семантичний, аналіз мовлення, синтаксичний та прагматичний аналізи.

Рис. 2.1. Етапи NLP [24]

Морфологічний аналіз у рамках обробки природної мови є одним із фундаментальних етапів аналізу тексту, який передбачає вивчення внутрішньої структури слів. Цей процес включає розпізнавання морфем, таких як корені, префікси, суфікси, а також визначення граматичних характеристик слів, таких як частини мови, відмінки, числа, роди, часи та інші граматичні категорії.

Процес морфологічного аналізу передбачає використання кількох підходів і технологій:

- *Правила морфології:*

використання набору правил, що описують можливі морфологічні зміни слів у мові. Це можуть бути правила відмінювання, дієвідмінювання, утворення ступенів порівняння тощо.

- *Морфологічні словники:*

використання словників, що містять інформацію про морфологічну структуру слів, їхні форми та граматичні характеристики.

- *Статистичні методи та машинне навчання:*

використання алгоритмів машинного навчання для прогнозування морфологічних характеристик слів на основі великих обсягів текстових даних.

Синтаксичний аналіз – процес визначення граматичної структури речення. Він включає в себе виявлення зв'язків між словами та побудову граматичних структур, таких як фрази і речення, що відповідають правилам граматики певної мови. Синтаксичний аналіз проводиться для розуміння сенсу

тексту, оскільки він дозволяє комп'ютерам визначати, як слова взаємодіють одне з одним у реченні [25].

Синтаксичний аналіз може здійснюватися за допомогою кількох підходів:

- *Правила контекстно-вільної граматики (CFG):*

використання набору граматичних правил, що визначають, як фрази можуть бути зібрані з окремих слів. Це дозволяє створювати синтаксичні дерева на основі визначених правил.

- *Моделі на основі залежностей:*

використання моделей, що будують дерево залежностей, де кожне слово в реченні пов'язане з іншими словами за допомогою певних граматичних відношень.

- *Статистичні методи та машинне навчання:*

використання алгоритмів машинного навчання для прогнозування синтаксичних структур на основі великих обсягів анотованих даних. Це може включати методи, такі як приховані марківські моделі (HMM), рекурентні нейронні мережі (RNN) і трансформери.

Синтаксичний аналіз є важливим для багатьох застосувань NLP, таких як машинний переклад, автоматичне резюмування текстів, пошук інформації, розуміння природної мови та інші. Він дозволяє зменшити неоднозначність і підвищити точність наступних етапів обробки тексту, таких як семантичний аналіз.

Семантичний аналіз у контексті обробки природної мови є процесом визначення і інтерпретації значення слів, фраз і речень у тексті. Цей етап обробки тексту спрямований на те, щоб комп'ютери могли зрозуміти не лише структуру тексту, а й його зміст, тобто сенс, що стоїть за словами. Семантичний аналіз є важливим для точного розуміння контексту, намірів та значення тексту [26].

Основні завдання семантичного аналізу включають:

- *Лексична семантика:*

визначення значення окремих слів у контексті. Це може включати розв'язання багатозначності слів, тобто визначення конкретного значення слова залежно від контексту.

- *Аналіз смислових відношень:*

визначення відношень між словами в тексті, таких як синонімія, антонімія, гіперонімія і гіпонімія. Це допомагає краще зрозуміти зв'язки між різними елементами тексту.

- *Розбір пропозицій:*

аналіз структури і значення речень для визначення того, хто що робить, коли і де. Це включає визначення ролей учасників дії (агент, пацієнт, інструмент тощо) і відношень між ними.

- *Аналіз намірів і контексту:*

визначення намірів і мети автора тексту, а також розуміння контексту, в якому були написані або сказані слова. Це важливо для виявлення прихованих смислів, сарказму або іронії.

- *Виділення сутностей і відношень:*

ідентифікація і класифікація ключових об'єктів у тексті, таких як імена, дати, місця, організації та інші сутності, а також визначення відношень між ними. Це дозволяє структурувати інформацію і полегшує її подальший аналіз.

Аналіз мовлення, або обробка мовлення (Speech Processing), є підгалуззю обробки природної мови і штучного інтелекту, яка займається перетворенням, аналізом і розумінням мовленнєвих сигналів. Цей процес включає декілька етапів, від розпізнавання мови до синтезу мовлення, і має важливе значення для створення систем, які можуть взаємодіяти з людьми на природному мовному рівні. Одним із основних завдань аналізу мовлення є розпізнавання мови, яке передбачає перетворення мовленнєвих сигналів у текст (Automatic Speech Recognition, ASR). Це важливий крок для додатків, таких як голосові помічники, диктування тексту та субтитрування в реальному часі [27]. Використання акустичних моделей для інтерпретації звуків і мовних моделей

для передбачення ймовірності послідовності слів є ключовим компонентом цього процесу.

Ще однією важливою складовою аналізу мовлення є синтез мовлення, який перетворює текст у природне мовлення (Text-to-Speech, TTS). Це дозволяє створювати системи, які можуть "говорити", що є корисним для створення озвучених відповідей у голосових помічниках, навігаційних системах та інших додатках. Синтез мовлення вимагає моделей, які здатні генерувати природне звучання на основі тексту, з урахуванням інтонації, ритму та акценту.

Аналіз мовлення також включає виявлення і розпізнавання мовця, що дозволяє системам визначати, хто говорить, і адаптуватися до специфічних особливостей їхньої мови. Це важливо для безпеки, персоналізації та покращення точності розпізнавання мови. Наприклад, системи безпеки можуть використовувати біометричні дані голосу для ідентифікації користувачів.

Прагматичний аналіз у контексті NLP є процесом розуміння значення висловлювань у контексті їх використання, тобто визначення того, що саме мовці мають на увазі у конкретній ситуації. Прагматика фокусується на тому, як контекст впливає на інтерпретацію мовленнєвих актів, включаючи мовленнєві наміри, імплікації та міжособистісні взаємодії [28].

Прагматичний аналіз враховує контекст, у якому було зроблено висловлювання, включаючи інформацію про мовця, слухача, попередні висловлювання та ситуацію в цілому. Це підкреслює важливість контекстуального розуміння для правильної інтерпретації висловлювань. Аналіз мовленнєвих актів розглядає висловлювання як дії, що виконуються за допомогою мови, такі як твердження, запитання, накази, прохання, обіцянки тощо. Прагматичний аналіз визначає, яку дію виконує мовець своїм висловлюванням, що допомагає зрозуміти його наміри і цілі.

Прагматичний аналіз також включає виявлення імплікатур, тобто прихованих значень, які не виражені явно, але зрозумілі з контексту.

Врахування контексту, намірів мовця та імплікатур допомагає досягти глибшого розуміння висловлювань і покращити взаємодію між людьми та

комп'ютерними системами. Прагматичний аналіз відіграє важливу роль у багатьох застосуваннях NLP, включаючи чат-боти, голосові помічники та системи підтримки клієнтів, де правильна інтерпретація намірів користувачів є критично важливою для надання відповідних відповідей і послуг [29].

Виявлення фейків за допомогою обробки природної мови (NLP) включає кілька ключових етапів, починаючи від збору даних і закінчуючи впровадженням моделей у реальні системи. Спершу, потрібно зібрати дані, які будуть використовуватися для навчання та тестування моделей. Це можуть бути тексти з різних джерел, включаючи новинні статті, відгуки, публікації в соціальних мережах і коментарі. Важливо зібрати як автентичні, так і підроблені тексти для створення збалансованого набору даних.

Після збору даних їх потрібно підготувати для аналізу. Це включає кілька кроків: токенізацію, що передбачає розбиття тексту на окремі слова або токени; видалення стоп-слів, тобто загальних слів, які не несуть смислового навантаження; стемінг і лемматизацію, що зводить слова до їх базової або кореневої форми; нормалізацію, яка приводить текст до стандартного формату, включаючи приведення до нижнього регістру та видалення пунктуації.

Наступним етапом є екстракція ознак з тексту, які будуть використовуватися для навчання моделей. Це можуть бути лексичні ознаки, такі як частота слів, біграми і триграми; стилістичні ознаки, такі як довжина речень і складність синтаксичних структур; семантичні ознаки, такі як векторні уявлення слів, наприклад, Word2Vec або GloVe; і соціальні ознаки, такі як метадані, включаючи час публікації, кількість лайків і коментарів, а також походження контенту.

На цьому етапі використовуються методи машинного навчання для навчання моделей на основі витягнутих ознак. Популярні підходи включають методи на основі правил, традиційні алгоритми машинного навчання, такі як логістична регресія, дерева рішень і метод опорних векторів (SVM), а також глибоке навчання, яке використовує нейронні мережі, такі як рекурентні нейронні мережі (RNN) або трансформери, як-от BERT чи GPT [30].

Після навчання моделі потрібно оцінити її ефективність на тестових даних. Використовуються метрики, такі як точність, повнота, F1-міра та ROC-AUC, що дозволяє визначити, наскільки добре модель розпізнає підробки. Якщо модель показала хороші результати на етапі оцінки, її можна впроваджувати в практичне використання. Це може включати інтеграцію в системи моніторингу новин, платформи соціальних мереж, сервіси електронної комерції тощо.

Після впровадження моделі необхідно постійно моніторити її ефективність і вдосконалювати на основі нових даних. Це може включати регулярне оновлення набору даних, перенавчання моделі і коригування параметрів. Отже, практичний процес виявлення підробок за допомогою NLP включає кілька етапів, від збору даних до впровадження моделей у реальні системи, з постійним моніторингом і вдосконаленням для забезпечення високої точності та ефективності.

2.2 Використання розумного аналізу даних

Термін «Інтелектуальний аналіз даних» має свої корені у понятті Data Mining. Цей термін складається з двох частин: «Data» (дані) і «Mining» (видобуток), що підкреслює суть процесу - пошук цінної інформації у великих обсягах даних. Data Mining передбачає або просіювання величезної кількості сирого матеріалу з метою виявлення корисних відомостей, або розумне дослідження і виявлення цінностей у даних.

В українській мові термін Data Mining часто перекладають як видобуток даних, витягування інформації, розкопування даних, інтелектуальний аналіз даних, засоби пошуку закономірностей, вилучення знань, аналіз шаблонів, або розкопування знань у базах даних. Незалежно від конкретного перекладу, всі ці варіанти підкреслюють ідею того, що Data Mining є складним процесом виявлення корисних закономірностей, зв'язків і знань із великих масивів даних. Це може включати використання різних алгоритмів, методів машинного

навчання та статистичних технік для того, щоб відкрити приховані інсайти та перетворити великі набори даних на цінну інформацію, яка може бути використана для прийняття обґрунтованих рішень у різних галузях, від бізнесу до науки і техніки [31].

Загальний процес виявлення фейків за допомогою розумного аналізу даних передбачає два етапи: виділення ознак та побудова моделі.

Етап виділення ознак спрямований на перетворення новинного контенту та пов'язаної з ним допоміжної інформації у формальні математичні структури. Це дозволяє моделям машинного навчання ефективніше розрізняти фейкові та справжні новини, ґрунтуючись на представлених ознаках. Важливо зазначити, що процес виділення інформації з традиційних медіа та соціальних мереж відрізняється. Традиційні ЗМІ покладаються на контекст новин, тоді як соціальні мережі можуть надавати додаткову допоміжну інформацію для виявлення фейкових новин. Тому контекст новин і соціальний контекст розглядаються окремо.

Джерело новини, автор або видавець, є важливим аспектом для оцінки достовірності інформації. Короткий заголовок привертає увагу читачів і описує основну тему статті, тоді як основний текст розкриває деталі новини, часто містить основне твердження, яке формує точку зору видавця. Візуальний контент, зокрема зображення та відео, надає додаткові підказки та допомагає у формуванні сприйняття новини. Новинний контент здебільшого складається з лінгвістичних і візуальних елементів.

Фейкові новини, створені з метою фінансової або політичної вигоди, часто використовують упереджену та провокаційну мову. Вони мають на меті привернути увагу за допомогою сенсаційних заголовків, що робить їх відмінними від справжніх новин. Для виявлення фейкових новин використовуються як загальні лінгвістичні ознаки, так і специфічні ознаки, характерні для певної тематики. До загальних лінгвістичних ознак відносяться лексичні особливості, такі як загальна кількість слів, кількість символів у слові, частота вживання великих слів та унікальних слів, а також синтаксичні

особливості, включаючи частоту службових слів, словосполучень, пунктуацію та маркування частин мови [32]. Специфічні лінгвістичні ознаки можуть включати цитовані слова, використання зовнішніх посилань, кількість графів і середню довжину графів. Додатково, програмне забезпечення для виявлення брехні може створювати додаткові функції для ідентифікації неправдивих сигналів.

Фейкові новини часто використовують фейкові зображення та відео для емоційного впливу на аудиторію. Візуальні ознаки, витягнуті з цих елементів, допомагають у виявленні фейкових новин. Оцінка візуальних ознак включає аналіз якості зображень, узгодженості візуального контенту, схожості між зображеннями, різноманітності візуальних елементів та їх кластеризації. Статистичні ознаки включають кількісні параметри візуального контенту, такі як загальна кількість зображень, співвідношення різних типів зображень, співвідношення кількох зображень, співвідношення гарячих зображень та співвідношення довгих зображень.

Ознаки соціального контексту можна отримати з соціальної активності користувачів при споживанні новин на платформах соціальних мереж. Існують три основні аспекти соціального контексту, які ми хочемо представити: користувачі, згенеровані пости та мережі. Далі ми розглянемо, як можна виокремити та представити особливості соціального контексту з цих трьох аспектів, щоб допомогти у виявленні фейкових новин.

По-перше, аналізуючи ознаки на основі користувачів, можна виявити, що фейкові новини часто створюються і поширюються акаунтами, які не є реальними людьми, наприклад, соціальними ботами або кіборгами. Фіксація профілів і характеристик користувачів за допомогою користувацьких функцій може надати корисну інформацію для виявлення фейкових новин. Ці ознаки можна класифікувати на різних рівнях: індивідуальному та груповому. Характеристики індивідуального рівня визначають достовірність і надійність користувача, використовуючи такі дані, як дата реєстрації, кількість підписників і твітів, тоді як характеристики групового рівня відображають

загальні характеристики користувачів, пов'язані з новинами. Наприклад, користувачі, які поширюють фейкові та справжні новини, можуть походити з різних спільнот з унікальними характеристиками, що відображаються в групових ознаках. Показники на рівні груп агрегують індивідуальні характеристики, такі як відсоток перевірених користувачів і середня кількість підписників, що дозволяє проводити точніший аналіз [33].

По-друге, ознаки на основі дописів мають важливе значення для виявлення фейкових новин у соціальних мережах, оскільки люди висловлюють свої емоції та думки щодо новин за допомогою дописів, якими вони діляться. Ці ознаки можна класифікувати як пост-рівень, груповий рівень і часовий рівень. Характеристики на рівні допису використовуються для створення значень характеристик для кожного допису, застосовуючи лінгвістичні підходи та підходи до вбудовування в новинний контент. Ці ознаки відображають соціальні реакції громадськості, такі як позиція, тема і достовірність. Показники позиції вказують на думку користувачів щодо новини, тоді як показники теми можуть бути вилучені за допомогою моделей тем, таких як латентний розподіл Діріхле (LDA). Функції достовірності оцінюють ступінь надійності, тоді як функції групового рівня агрегують значення характеристик для відповідних постів, використовуючи "мудрість натовпу". Функції часового рівня враховують часові варіації значень характеристик на рівні постів, а неконтрольовані методи вбудовування, такі як рекурентна нейронна мережа (RNN), фіксують зміни в постах у часі. Математичні характеристики, такі як параметри SpikeM, можна обчислити на основі форми часового ряду для відповідних метрик.

По-третє, процеси поширення фейкових новин мають тенденцію формувати цикл ехо-камери, що підкреслює цінність вилучення мережевих ознак для представлення цих типів мережевих патернів для виявлення фейкових новин. Мережеві ознаки виділяються шляхом побудови певних мереж серед користувачів, які опублікували відповідні пости в соціальних мережах. Можна побудувати різні типи мереж, зокрема мережі позицій, мережі

спільних публікацій і мережі дружби. Мережа позицій може бути побудована з вузлами, що вказують на всі твіти, які мають відношення до новини, і ребрами, що вказують на вагу схожості позицій. Мережа збігів підраховує активність користувачів, щоб побачити, чи писали ці користувачі схожі пости щодо одних і тих самих новинних статей. Мережа дружби представляє структуру фоловерів/фоловерів користувачів, які публікують пов'язані твіти. Мережа дифузії, яка є розширенням мережі дружби, відстежує поширення новин, де вузли представляють користувачів, а ребра - шляхи поширення інформації. Наприклад, шлях поширення інформації між двома користувачами існує тоді і тільки тоді, коли один користувач слідкує за іншим і публікує повідомлення про певну новину тільки після того, як це зробить інший користувач. Існуючі мережеві метрики можуть бути використані як представлення ознак після належної побудови, наприклад, коефіцієнти ступеня та кластеризації для мереж дифузії та дружби. Інші підходи вивчають латентні особливості вбудовування вузлів за допомогою SVD або алгоритмів мережевого поширення [34].

Другим етапом іде побудова моделі. У цьому етапі розглядаються деталі процесу побудови моделей для декількох існуючих підходів, які класифікуються на основі їхніх основних джерел вхідних даних: моделі новинного контенту та моделі соціального контексту.

Моделі новинного контенту фокусуються на підходах, заснованих на знаннях і стилі, для класифікації фейкових новин. Підходи, засновані на знаннях, використовують зовнішні джерела для перевірки фактів у новинному контенті, присвоюючи значення правдивості конкретному контексту. Така перевірка фактів може бути експертною, краудсорсинговою або обчислювальною.

Експертно-орієнтована перевірка фактів покладається на людей-експертів для верифікації даних і документів, як це робиться на таких веб-сайтах, як PolitiFact і Snopes. Проте цей метод є інтелектуально складним і трудомістким, що обмежує його ефективність і масштабованість. Фактчекінг, орієнтований на краудсорсинг, використовує "мудрість натовпу", дозволяючи користувачам

коментувати новинний контент і агрегувати їхню оцінку достовірності. Прикладами таких платформ є Fiskkit і бот для боротьби з фейковими новинами "For real" від LINE, який дозволяє користувачам повідомляти про підозрілий контент, який потім перевіряється редакторами.

Комп'ютерно-орієнтована перевірка фактів спрямована на створення автоматичної, масштабованої системи класифікації правдивих і неправдивих тверджень. Вона вирішує дві основні задачі: виявлення тверджень, що заслуговують на перевірку, і визначення їхньої правдивості. Для виявлення таких тверджень виокремлюються фактичні твердження в новинному контенті, що передають ключові ідеї та точки зору. Перевірка фактів спирається на зовнішні ресурси, такі як відкриті веб-джерела та структурований граф знань. Відкриті веб-джерела надають часті посилання для порівняння тверджень, а графи знань інтегрують пов'язані дані, такі як DBpedia та Google Relation Extraction Corpus, для перевірки можливості виведення тверджень із наявних фактів.

Підходи на основі стилю фокусуються на виявленні маніпуляцій у стилі написання контенту. Видавці фейкових новин використовують специфічні стилі написання, щоб апелювати до широкого кола споживачів. Підходи, орієнтовані на обман, фіксують оманливі заяви в новинному контенті, використовуючи вдосконалені моделі обробки природної мови, такі як глибокий синтаксис і риторична структура. Глибокий синтаксис застосовує імовірнісну контекстну граматику (PCFG) для перетворення речень на правила, що дозволяють виявляти фрази обману. Теорія риторичної структури фіксує відмінності між оманливими та правдивими реченнями. Для класифікації правдивості новин також використовуються глибинні мережеві моделі, такі як CNN.

Підходи, орієнтовані на об'єктивність, виявляють стилістичні сигнали, які вказують на зниження об'єктивності в новинному контенті. Гіперпартійні стилі, що надають перевагу певній політичній партії, виявляються за допомогою лінгвістичних особливостей. Жовта журналістика використовує заголовки, що

привертають увагу, та перебільшення, часто застосовуючи клікбейти для нагнітання страху. Ці оманливі заголовки слугують індикаторами для розпізнавання фейкових новин.

Моделі соціального контексту використовують дані з соціальних медіа для вдосконалення моделей новинного контенту, залучаючи користувачів з різними точками зору. Існуючі підходи до виявлення фейкових новин у соціальних медіа поділяються на дві основні категорії: підходи, засновані на позиціях, і підходи, засновані на поширенні інформації. Хоча ці підходи ще не набули широкого застосування для виявлення фейкових новин, вони успішно використовуються для виявлення чуток і мають великий потенціал у боротьбі з фейковими новинами [35].

Підходи, засновані на позиціях, використовують погляди користувачів, виражені у відповідних дописах, щоб визначити правдивість новинних статей. Явна позиція включає прямі вираження емоцій або думок, такі як реакції "великий палець вгору" або "великий палець вниз". Неявна позиція, навпаки, визначається через аналіз текстів постів у соціальних мережах. Завдання виявлення позицій полягає в автоматичному визначенні прихильності, нейтральності чи неприхильності користувача до певного суб'єкта, події або ідеї. Раніші методи класифікації позицій здебільшого покладалися на ручну лінгвістичну обробку або вбудовування ознак в окремі пости. Сучасні методи, такі як тематичне моделювання за допомогою латентного розподілу Діріхле (LDA), дозволяють вивчати приховані позиції, аналізуючи теми постів і роблячи висновки про правдивість новин на основі цих позицій.

Підходи до виявлення фейкових новин, засновані на поширенні інформації, спираються на взаємозв'язки між постами у соціальних мережах для прогнозування достовірності новин. Основна ідея полягає в тому, що достовірність новинної події тісно пов'язана з достовірністю постів, що про неї пишуть. Для аналізу процесів поширення новин можна створювати як гомогенні, так і гетерогенні мережі довіри. Гомогенні мережі складаються з

одного типу об'єктів, наприклад, лише постів або подій. Гетерогенні мережі включають різні типи об'єктів, такі як пости, підподії та основні події.

2.3 Використання нейронних мереж

Використання нейронних мереж для виявлення підробок стало передовим і ефективним методом завдяки потужним можливостям цих моделей у розпізнаванні шаблонів і виявленні відхилень від норми.

Для виявлення підробок зображень згорткові нейронні мережі (CNN) виявилися особливо ефективними завдяки своїй здатності обробляти та аналізувати візуальні дані. Структура CNN складається з декількох шарів, включаючи згорткові шари, об'єднані шари та повністю з'єднані шари. Згорткові шари діють як екстрактори ознак, скануючи вхідні зображення для виявлення країв, текстур і візерунків. Об'єднані шари зменшують розмірність цих ознак, зберігаючи важливу інформацію і роблячи обчислення більш ефективними. Повністю з'єднані шари, в свою чергу, інтерпретують ці особливості для класифікації зображень або виявлення специфічних об'єктів і аномалій.

Ефективність CNN у виявленні підробок зображень зумовлена їхньою здатністю навчатися просторовим ієрархіям ознак на великій кількості зображень. Під час навчання CNN отримують великі набори даних, що містять як справжні, так і підроблені зображення. Цей процес дозволяє мережі навчитися відрізняти справжні зображення від підроблених. Наприклад, виявляючи маніпуляції з фотошопом, CNN може виявити невідповідності в освітленні, тінях і текстурах, які можуть бути не одразу помітні людському оку. Аналогічно, при виявленні склеєних зображень, CNN можуть виявити невідповідність меж і різну роздільну здатність, що вказує на те, що частини зображення були штучно об'єднані.

Вдосконалені архітектури CNN, такі як ResNet, Inception і EfficientNet, ще більше розширюють можливості виявлення підробок, впроваджуючи більш

складні механізми для вилучення та інтеграції ознак. Ці мережі можуть обробляти глибші та складніші моделі, підвищуючи точність виявлення дрібних і складних підробок. Крім того, може застосовуватися трансферне навчання, коли попередньо навчені моделі на великих наборах зображень підлаштовуються під конкретні завдання виявлення підробок, що значно скорочує необхідний час і ресурси для навчання, одночасно підвищуючи продуктивність.

На практиці системи на основі CNN використовуються в різних галузях, включаючи цифрову криміналістику, автентифікацію медіа та безпеку. Вони є важливими інструментами для перевірки автентичності зображень у журналістиці, судових доказах і контенті соціальних мереж. Оскільки витонченість підробок зображень продовжує зростати, системи CNN та їхні вдосконалені варіанти залишаються на передовій технологічного захисту від підробки цифрових зображень, забезпечуючи цілісність і достовірність візуальної інформації в цифрову епоху [36].

При виявленні фальсифікацій відео нейронні мережі розширюють свої можливості до послідовностей кадрів, що дозволяє їм аналізувати не тільки просторову інформацію в кожному кадрі, але й часову динаміку між кадрами. Це має вирішальне значення для виявлення підробок відео, які часто пов'язані з тонкими маніпуляціями, які може бути складно виявити, розглядаючи окремі кадри ізольовано.

Рекурентні нейронні мережі (RNN) та їхні вдосконалені варіанти, мережі з довгою короткочасною пам'яттю (LSTM), є особливо ефективними в цьому контексті. RNN призначені для обробки послідовних даних, зберігаючи пам'ять про попередні входи, що дозволяє їм фіксувати часові залежності. LSTM розширюють ці можливості, використовуючи вентилі для керування потоком інформації, ефективно керуючи довготривалими залежностями та пом'якшуючи проблему зникаючого градієнта, яка може перешкоджати роботі стандартних RNN.

У поєднанні зі згортковими нейронними мережами (CNN) цей комбінований підхід використовує сильні сторони обох архітектур. CNN спочатку обробляють кожен кадр, щоб виділити просторові особливості, такі як краї, текстури та об'єкти. Потім ці особливості передаються до RNN або LSTM, які аналізують послідовність кадрів для виявлення часових невідповідностей та аномалій. Цей метод особливо ефективний для виявлення складних відеопідробок, таких як "глибокі фейки", коли штучний інтелект використовується для створення високореалістичних фальшивих відео, змінюючи вираз обличчя, рухи губ або навіть цілі образи на відеозаписах.

Deepfakes становлять значну проблему через їхній високий рівень реалістичності, який може ввести в оману як людей-спостерігачів, так і традиційні методи виявлення. Аналізуючи часову узгодженість виразів обличчя, рухів очей та інших тонких сигналів у кадрах, нейронні мережі можуть виявити невідповідності, які вказують на маніпуляції. Наприклад, невідповідність у морганнях або неприродні переходи між кадрами можуть свідчити про процеси глибокої генерації фейків [37].

Удосконалені архітектури нейронних мереж, такі як 3D CNN і моделі на основі трансформаторів, ще більше підвищують ефективність виявлення підробок відео завдяки більш ефективному захопленню просторової та часової інформації. Ці моделі можуть обробляти все відео як тривимірну структуру, враховуючи часовий вимір поряд з висотою і шириною кадрів, що покращує їхню здатність виявляти аномалії, які охоплюють кілька кадрів.

У практичному застосуванні ця технологія має важливе значення для забезпечення цілісності відеоконтенту в таких сферах, як новинні ЗМІ, судові докази та соціальні мережі. Оскільки підробки відео стають все більш витонченими, інтеграція CNN з RNN або LSTM, разом з новітніми архітектурами, забезпечує надійну основу для виявлення і пом'якшення впливу фальшивих відео, підтримуючи довіру і автентичність цифрових медіа.

Для виявлення аудіопідробки нейронні мережі, такі як згорткові нейронні мережі (CNN) та мережі з довгою короткочасною пам'яттю (LSTM), відіграють

вирішальну роль в аналізі аудіосигналів. Ці моделі вміють виявляти тонкі розбіжності в різних звукових характеристиках, які можуть вказувати на те, чи є запис автентичним, чи сфабрикованим. Тренуючи ці мережі на наборах даних, що містять як справжні, так і підроблені аудіо-зразки, вони вчаться розрізняти їх на основі особливостей мови, фонових шумів та інших релевантних характеристик.

Одним з ефективних методів у цій галузі є аналіз спектрограм, де аудіосигнали перетворюються на візуальні зображення, які відображають частотний спектр звуку в часі. Спектрограми є багатим джерелом інформації для нейронних мереж, оскільки вони фіксують складні деталі звукових хвиль, які можуть бути пропущені в необробленому аудіосигналі. ШНМ, які чудово справляються із завданнями обробки зображень, можуть бути застосовані до цих спектрограм, щоб виявити аномалії та патерни, які вказують на підробку.

Наприклад, за допомогою CNN можна виявити неприродні зсуви висоти або тону, невідповідності у фоновому шумі або порушення в гармонійній структурі звуку, які вказують на маніпуляції. З іншого боку, LSTM добре підходять для виявлення часових залежностей в аудіоданих. Аналізуючи послідовності аудіокадрів, LSTM можуть виявити часові невідповідності, такі як неприродні паузи або раптові зміни звукових характеристик, які часто зустрічаються у підроблених записах [38].

Поєднання CNN і LSTM дозволяє проводити комплексний аналіз як просторових особливостей спектрограми, так і часової динаміки аудіосигналу. Такий гібридний підхід підвищує здатність виявляти підробки, використовуючи сильні сторони обох типів нейронних мереж.

Вдосконалені архітектури та методи нейронних мереж, такі як вейвлет-перетворення та рекурентні згорткові структури, ще більше покращують можливості виявлення, забезпечуючи більш детальний та контекстно-орієнтований аналіз аудіосигналів. Ці моделі можуть аналізувати аудіо в різних масштабах і роздільній здатності, захоплюючи як високочастотні деталі, так і ширші часові патерни.

У практичному застосуванні системи виявлення підробок аудіо є критично важливими в різних галузях, включаючи криміналістичний аналіз, перевірку медіа та безпеку. Вони допомагають забезпечити цілісність аудіозаписів, що використовуються в судовому процесі, перевірити автентичність трансльованих медіа та виявити шахрайські дії в телекомунікаційних системах. Оскільки підробки аудіозаписів стають все більш витонченими, постійний розвиток методів виявлення на основі нейронних мереж має важливе значення для збереження достовірності та надійності аудіоінформації в цифрову епоху.

Виявлення підробки тексту використовує передові моделі нейронних мереж, зокрема архітектури на основі трансформаторів, такі як BERT (Bidirectional Encoder Representations from Transformers) і GPT (Generative Pre-trained Transformer), які добре розуміють і генерують людську мову. Ці моделі чудово справляються із завданнями обробки природної мови завдяки здатності фіксувати контекстуальні зв'язки в тексті за допомогою механізмів уваги, що дозволяє їм виявляти тонкі шаблони та невідповідності, які можуть свідчити про підробку.

BERT і GPT можна навчати на великих масивах даних, що містять як автентичні, так і підроблені тексти. Таке навчання дозволяє їм вивчати нюанси справжнього письма, зокрема граматичні структури, стилістичні патерни та семантичну зв'язність. Аналізуючи текст на автентичність, ці моделі можуть виявляти синтаксичні аномалії, такі як незвичний порядок слів або граматичні помилки, що відхиляються від типового вживання. Крім того, вони можуть виявляти семантичні невідповідності, коли зміст тексту є непослідовним або нелогічним, що часто є ознаками підробки.

Наприклад, BERT, завдяки двонаправленому підходу, читає текстові послідовності в обох напрямках (зліва направо і справа наліво), що дозволяє йому більш повно зрозуміти контекст кожного слова. Ця здатність робить BERT особливо ефективним для виявлення невідповідностей у реченні або між кількома реченнями. GPT, який чудово генерує зв'язний і контекстуально

відповідний текст, можна використовувати для порівняння згенерованого тексту з вхідним текстом, виділяючи розбіжності, які свідчать про маніпуляції [39].

Ці моделі також отримують вигоду від трансферного навчання, коли їх попередньо тренують на великих масивах тексту з різних джерел, а потім допрацьовують на конкретних завданнях, пов'язаних із виявленням підробок. Цей процес покращує їхню здатність узагальнювати бачені приклади до нових, небачених текстових підробок.

На практиці моделі на основі Transformer можна застосовувати для різних форм виявлення підробки тексту, зокрема для виявлення плагіату, ідентифікації фейкових новин та перевірки автентичності документів. Вони можуть виявити, коли текст було змінено, порівнюючи його з базовим відомим автентичним текстом, виділяючи відхилення, які можуть бути невидимими для людей-рецензентів. Крім того, вони здатні виявляти підробки, створені іншими системами штучного інтелекту, які розроблені для імітації людського стилю письма.

Постійний розвиток цих моделей у поєднанні з їхньою здатністю аналізувати великі масиви даних і розпізнавати складні патерни гарантує, що вони залишаються на передовій у виявленні підробок текстів. Підтримуючи цілісність письмового контенту на цифрових платформах, ці технології допомагають зберегти достовірність і надійність інформації у світі, який дедалі більше залежить від тексту.

Висновки до розділу 2

Розділ представляє детальний аналіз сучасних підходів до ідентифікації неправдивої інформації, що поширюється через текстові джерела. Розглядаються три основні методи: обробка природної мови, розумний аналіз даних та використання нейронних мереж.

Метод обробки природної мови (NLP) є одним із найбільш потужних інструментів у виявленні фейкових новин. Цей метод дозволяє аналізувати текстові дані на рівні синтаксису, семантики та прагматики. Використання NLP включає такі техніки, як токенізація, лемматизація, частеречова розмітка та векторизація тексту. Ці процеси дозволяють перетворити текст у формат, придатний для подальшого аналізу за допомогою машинного навчання. Дослідження показують, що застосування NLP для виявлення фейків може включати аналіз стилістичних особливостей тексту, виявлення аномалій у вживанні слів та фраз, а також ідентифікацію емоційного забарвлення тексту, яке часто використовують у фейкових новинах для маніпулювання аудиторією.

Використання розумного аналізу даних спрямоване на обробку великих обсягів текстового контенту з метою виявлення закономірностей, які можуть вказувати на фейковий характер інформації. Цей метод включає збір, очищення та обробку даних з різних джерел, таких як новинні портали, соціальні мережі та блоги. Аналітика великих даних дозволяє виявити тенденції в поширенні фейкових новин, включаючи аналіз їхнього впливу на громадську думку та визначення ключових платформ для їхнього розповсюдження. Використання методів кластерного аналізу, класифікації та регресії дозволяє створювати моделі, які можуть передбачати ймовірність того, що певний текстовий контент є фейковим.

Використання нейронних мереж, особливо глибоких нейронних мереж, є ще одним потужним підходом до виявлення фейків у текстовому контенті. Нейронні мережі здатні автоматично навчатися на великих наборах даних і виявляти складні патерни, характерні для фейкових новин. Однією з популярних архітектур є рекурентні нейронні мережі (RNN) та їх варіації, такі як довготривала короткочасна пам'ять (LSTM) та трансформери. Ці моделі здатні аналізувати текст у контексті, враховуючи послідовність слів та взаємозв'язок між ними. Дослідження показують, що нейронні мережі можуть досягати високої точності у класифікації текстів на правдиві та неправдиві, забезпечуючи при цьому високу швидкість обробки даних.

Таким чином, розділ демонструє комплексний підхід до боротьби з дезінформацією, підкреслюючи важливість обробки природної мови, розумного аналізу даних та використання нейронних мереж. Кожен з цих методів робить вагомий внесок у загальну стратегію ідентифікації та протидії фейковим новинам, забезпечуючи різноманітні інструменти та техніки для підвищення точності та ефективності виявлення неправдивої інформації.

РОЗДІЛ 3 РОЗРОБКА СИСТЕМИ МАШИННОГО НАВЧАННЯ

3.1. Порівняння та визначення найефективніших методів виявлення фейків

Серед різноманітних методів виявлення фейкових новин особливе місце належить підходам на основі глибокого навчання, серед яких особливу увагу заслуговує BERT. Цей метод використовує складну архітектуру трансформатора та механізми уваги для глибокого розуміння текстового контексту. Прагнучи досягти найвищої точності у різних завданнях обробки природної мови, BERT виявляється особливо ефективним у виявленні фейкових новин завдяки здатності аналізувати контекст слова в обох напрямках. Незважаючи на вражаючі переваги у високій точності та можливості адаптації до конкретних завдань, використання BERT вимагає значних обчислювальних ресурсів та об'ємних наборів даних, що ускладнює його впровадження. Такі вимоги можуть створити певні труднощі, особливо для організацій з обмеженими ресурсами. Однак в умовах правильної підготовки і достатнього фінансування BERT може стати потужним інструментом у боротьбі з поширенням дезінформації та фейкових новин, забезпечуючи високу ефективність та надійність у виявленні недостовірної інформації.

Гібридна архітектура нейронних мереж є одним з перспективних методів у сфері виявлення фейкових новин. Цей підхід об'єднує в собі дві потужні технології: згорткові нейронні мережі (CNN), що відповідають за вилучення ознак, і мережі довготривалої короткочасної пам'яті (LSTM), які відповідають за послідовну обробку даних. Ця комбінація дозволяє використовувати найкращі аспекти обох типів нейронних мереж, що призводить до покращення ефективності виявлення фейкових новин. Незважаючи на те, що гібридна архітектура проявляє себе як дієвий інструмент у вилученні ознак та обробці послідовних даних, варто відзначити, що її складність вимагає пристойного

налаштування та значних обчислювальних потужностей. Це може бути чинником, що ускладнює впровадження даного методу, особливо для організацій з обмеженими ресурсами. Однак при належній підготовці та відповідному фінансуванні гібридна архітектура нейронних мереж може стати важливим інструментом у боротьбі з поширенням дезінформації, забезпечуючи високу ефективність у виявленні фейкових новин і підвищуючи рівень надійності та об'єктивності інформаційних потоків [40].

Мультимодальне виявлення фейкових новин – стратегії, що включають як текстові, так і візуальні дані, такі як швидке навчання та методи злиття на основі схожості, пропонують цілісний підхід. Наприклад, методи, які використовують швидке навчання та злиття на основі схожості, надають цілісний підхід до виявлення фейкових новин. Ці стратегії дозволяють поєднувати різні джерела даних для отримання більш повного розуміння контенту, що в свою чергу підвищує точність виявлення. Однак, для успішної реалізації таких методів необхідно мати різноманітні набори даних, що містять як текстову, так і візуальну інформацію. Це призводить до збільшення складності процесу, оскільки вимагається обробка та аналіз різнорідних даних. Такі вимоги можуть ускладнити впровадження методів мультимодального виявлення фейкових новин, особливо для організацій з обмеженими ресурсами. Проте, при належній підготовці та відповідному фінансуванні, використання мультимодальних підходів може стати ефективним інструментом у боротьбі з поширенням дезінформації та підвищення рівня достовірності інформації.

Методи виділення ознак і адаптації до домену, зокрема вирівнювання підпросторів, виявляються перспективними стратегіями для покращення адаптивності моделей та їх узагальнення для різних наборів даних. Ця стратегія полягає у тому, що моделі навчаються вирівнювати простори ознак із різних доменів, що дозволяє їм краще адаптуватися до різних умов і контекстів, таким чином поліпшуючи їхню здатність до виявлення фейкових новин. Однак варто відзначити, що складність вирівнювання підпросторів та можлива втрата специфічних ознак, характерних для конкретного домену, створюють певні

виклики у впровадженні цього методу. Такі проблеми можуть виникнути через необхідність враховувати різноманітність даних і забезпечити адекватне представлення їх у моделі [41]. Попри це, за належної уваги та належних зусиль у напрямку налагодження цих методів, вони можуть стати корисними інструментами у виявленні фейкових новин, сприяючи підвищенню ефективності та надійності моделей.

Скомпонуємо проведений аналіз в Табл. 3.1.

Таблиця 3.1

Порівняльний аналіз методів виявлення фейків

Метод / Характеристика	BERT	CNN+LSTM	Мультимодальне виявлення фейкових новин	CLIP	Методи виділення ознак і адаптації до домену
Загальна характеристика	Використовує архітектуру трансформаторів та увагу для розуміння текстового контексту	Поєднання згорткових та LSTM мереж для аналізу ознак та послідовностей	Використання тексту та візуальної інформації для аналізу	Оцінка кореляції між текстом та зображеннями для виявлення невідповідностей	Узгодження просторів ознак для покращення перенесення моделей
Ефективність	Висока точність у виявленні фейкових новин	Покращена ефективність у виявленні фейкових новин	Підвищена точність завдяки комбінації даних з різних джерел	Ефективне виявлення фейкових новин, що містять текст та зображення	Підвищена точність виявлення та адаптивність моделей
Переваги	Надійне контекстне розуміння, висока точність і можливість налаштування	Ефективне вилучення ознак, обробка послідовних даних	Різноманітність даних, стійкість до різних типів фейкових новин	Можливість виявлення мультимедійного контенту	Узагальненість, адаптивність
Недоліки	Вимагає значних обчислювальних ресурсів	Складна архітектура, вимагає ретельного налаштування	Підвищена складність, потреба в різноманітних даних	Обмеженість у виявленні фейкових новин, вимагає високоякісних попередньо навчених моделей	Складність у узгодженні просторів ознак, можлива втрата специфічних ознак

Отже, вибір методу для виявлення фейкових новин неминуче залежить від конкретних вимог і обмежень, що виникають у зв'язку з цим завданням. Наприклад, враховуються наявність обчислювальних ресурсів, типи даних, які доступні для аналізу, і бажана точність результатів. Кожен метод має свої сильні та слабкі сторони, які слід ураховувати при виборі. Наприклад, методи, що базуються на глибокому навчанні, можуть забезпечити високу точність завдяки адаптивності до конкретного завдання, але вони вимагають значних обчислювальних ресурсів і великих обсягів даних для навчання. З іншого боку, методи, що використовують багатомодальні дані, можуть бути більш стійкими до різноманітних типів фейкових новин, але вони можуть потребувати складнішої обробки та аналізу даних. Гібридні підходи або поєднання декількох методів можуть запропонувати оптимальні результати, використовуючи переваги кожного підходу та компенсуючи їхні недоліки [42]. Такий гнучкий підхід дозволяє забезпечити ефективне виявлення фейкових новин у різноманітних умовах та обставинах.

3.2. Розробка системи виявлення фейків

Розробка системи виявлення фейкових новин - це складний та багатоетапний процес, який використовує різноманітні технології та методи для ефективного аналізу та класифікації інформації. Розглянемо процес розробки системи детальніше.

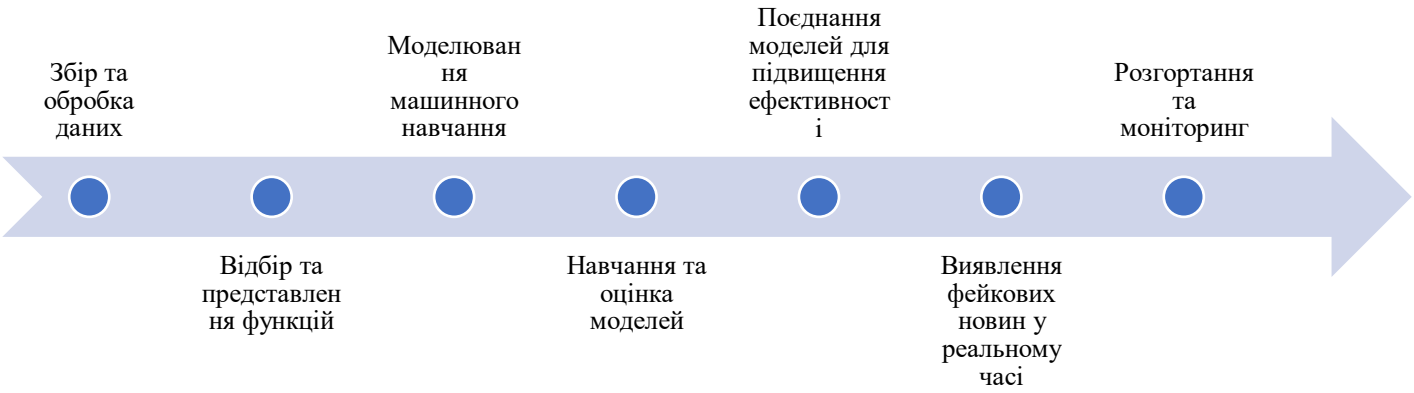

Рис. 3.1. План розробки системи виявлення фейкових новин

Щоб створити надійну систему виявлення підробки тексту, процес зазвичай починається з комплексного збору даних. Це передбачає впровадження веб-сканерів для збору новинних статей з широкого кола онлайн-ресурсів, включно з авторитетними новинними веб-сайтами та різноманітними платформами соціальних мереж. Для ефективного навчання моделі дуже важливо, щоб набір даних був збалансованим і містив різноманітний вибір як справжніх, так і фейкових новин.

Після збору даних наступним кроком є попередня обробка даних, яка включає кілька ключових завдань для підготовки текстових даних до аналізу. У таких мовах, як тайська або китайська, де немає пробілів між словами, застосовуються відповідні методи сегментації, щоб розбити текст на значущі лексеми або слова. Крім того, застосовуються методи очищення даних для видалення непотрібних символів, стоп-слів і будь-якого іншого шуму, який може заважати процесу аналізу. Це гарантує, що текстові дані будуть чистими і готовими до подальшої обробки [43].

Виділення ознак - ще один важливий аспект виявлення підробки тексту. Цей крок передбачає виявлення та вилучення з тексту ознак, які вказують на фейкові чи справжні новини. Найпоширеніші ознаки, які використовуються в цьому контексті, включають частоту термінів, зворотну частоту документів (TF-IDF) та оцінки настроїв. Ці ознаки допомагають виявити важливі аспекти тексту, які можна використати для розрізнення справжніх і сфабрикованих новинних статей.

На додаток до цих кроків, для подальшого підвищення точності виявлення підроблених текстів можна використовувати передові методи обробки природної мови, такі як моделі глибокого навчання, такі як BERT і GPT. Ці моделі чудово розуміють семантичне значення тексту і можуть виявляти тонкі закономірності та невідповідності, які можуть свідчити про підробку.

Поєднуючи комплексний збір даних, ретельну попередню обробку даних і передові методи вилучення ознак, а також найсучасніші моделі глибокого

навчання, системи виявлення підроблених текстів можуть досягти високого рівня точності та надійності. Ця технологія відіграє вирішальну роль у боротьбі з поширенням дезінформації та забезпеченні достовірності новин та інформації в цифрову епоху.

При виявленні підробки тексту вибір і представлення ознак є ключовими для ефективного розрізнення справжніх і сфабрикованих новинних статей. Для вилучення значущих ознак з текстових даних зазвичай використовують кілька методів:

- *Вбудовування в текст* слугують фундаментальним інструментом для перетворення текстових даних на числові вектори, які інкапсулюють семантичні зв'язки. Такі моделі вбудовування, як Word2Vec і GloVe, відображають слова зі словника в щільне векторне представлення в неперервному векторному просторі. Фіксуючи семантичну схожість і зв'язки між словами, ці вбудовування надають багату контекстну інформацію, яка покращує розуміння змісту тексту.
- *TF-IDF* (частота терміна - зворотна частота документа) - це ще одна важлива техніка представлення особливостей, яка використовується в аналізі тексту. Він кількісно оцінює важливість терміна в документі відносно його частоти в корпусі документів. Зважуючи терміни на основі їхньої частоти в документі та їхньої рідкісності в корпусі, TF-IDF виділяє терміни, які часто зустрічаються в документі та є унікальними для нього. Такий підхід допомагає виділити важливі ключові слова і фрази, які можуть бути показовими для змісту і теми документа.
- *Аналіз настроїв* додає ще один рівень складності до представлення особливостей, аналізуючи емоційний тон і полярність тексту. Оцінюючи настрої, виражені в новинних статтях, такі як позитивні, негативні або нейтральні, аналіз настроїв може зафіксувати суб'єктивні аспекти контенту. Ця інформація особливо корисна для виявлення фейкових новин, оскільки сфабриковані статті можуть демонструвати упереджені або перебільшені настрої, які відрізняються від справжніх новин.

Ці методи представлення особливостей доповнюють один одного і забезпечують різноманітні точки зору на текстові дані, що дає змогу проводити більш комплексний аналіз і виявляти фейкові новини. Використовуючи вставки слів, TF-IDF та аналіз настроїв, системи виявлення підроблених текстів можуть фіксувати як семантичне значення, так і контекстуальні нюанси, тим самим підвищуючи точність розрізнення справжніх і фейкових новинних статей [44].

У виявленні підробки тексту моделювання машинного навчання відіграє вирішальну роль в аналізі вилучених ознак і прогнозуванні автентичності новинних статей. Зазвичай використовуються дві широкі категорії моделей:

- *Традиційні моделі*: Традиційні моделі машинного навчання забезпечують міцну основу для завдань класифікації текстів.
- *Моделі глибокого навчання*: Вдосконалені моделі глибокого навчання пропонують найсучаснішу продуктивність у завданнях класифікації тексту.

Використовуючи як традиційні моделі, так і моделі глибокого навчання, системи виявлення підробки тексту можуть скористатися різноманітними підходами і методами. Традиційні моделі забезпечують інтерпретованість і ефективність, тоді як моделі глибокого навчання пропонують чудову продуктивність і здатність фіксувати складні патерни в текстових даних. Використовуючи сильні сторони кожного типу моделей, системи виявлення підробки тексту можуть досягти високого рівня точності та надійності в розрізненні справжніх і сфабрикованих новинних статей.

У процесі виявлення підробки тексту навчання та оцінка моделі є критично важливими етапами, які забезпечують ефективність і надійність розробленої системи. Ці етапи включають в себе:

- *Навчання*: Використання значної частини зібраних даних для навчання моделей машинного навчання. Важливо переконатися, що навчальний набір даних є різноманітним і репрезентативним для реального розподілу новинних статей. Методи перехресної перевірки, такі як k-кратна перехресна перевірка, зазвичай застосовуються для зменшення надмірного припасування та

гарантування того, що модель добре узагальнює невидимі дані. Розбиваючи набір даних на кілька підмножин, навчання виконується ітеративно на різних комбінаціях навчальних і перевірочних наборів, що дозволяє отримати більш надійну оцінку продуктивності моделі.

- *Метрики оцінки*: Оцінка продуктивності навчених моделей за допомогою відповідних оціночних метрик має вирішальне значення для визначення їхньої ефективності в розрізненні справжніх і сфабрикованих новинних статей. Найпоширеніші метрики оцінювання включають точність, достовірність, пригадування та показник F1. Точність вимірює частку правильно класифікованих примірників від загальної кількості примірників. Прецизійність - це відношення істинно позитивних прогнозів до загальної кількості позитивних прогнозів, що вказує на здатність моделі уникати помилкових спрацьовувань. Відтворення, також відоме як чутливість, вимірює частку істинних позитивних прогнозів від усіх фактично позитивних випадків. Показник F1 поєднує точність і відгук в одну метрику, забезпечуючи збалансовану оцінку ефективності моделі.

Використовуючи ці метрики оцінки, системи виявлення підробки тексту можуть ефективно кількісно оцінювати продуктивність навчених моделей і порівнювати різні підходи, щоб визначити найбільш підходяще рішення [45]. Крім того, такі методи, як налаштування гіперпараметрів і вибір моделі, можуть додатково оптимізувати роботу моделей, забезпечуючи їхню стійкість і надійність у реальних умовах.

Для підвищення продуктивності систем виявлення підробок тексту об'єднання декількох моделей за допомогою ансамблевих методів і гібридних архітектур може значно підвищити точність і надійність виявлення. Ці підходи включають в себе:

- *Ансамблеві методи*: Ансамблеві методи передбачають об'єднання прогнозів кількох окремих моделей для отримання більш точного і надійного остаточного прогнозу. Поширені ансамблеві методи включають класифікатори голосування, де прогнози різних моделей об'єднуються шляхом голосування, і

стекінг, де прогнози базових моделей використовуються як вхідні ознаки для метанавчального пристрою.

- *Гібридні моделі*: Гібридні моделі інтегрують різні типи архітектур нейронних мереж, щоб використовувати їхні сильні сторони.

Поєднуючи ансамблеві методи та гібридні архітектури, системи виявлення підроблених текстів можуть отримати вигоду від синергії між різними моделями та архітектурами, що призводить до підвищення продуктивності та надійності. Ці підходи дозволяють системі фіксувати ширший спектр ознак і закономірностей у даних, що призводить до більш точної ідентифікації справжніх і сфабрикованих новинних статей. Крім того, ансамблеві методи та гібридні архітектури надають вбудовані механізми для обробки невизначеності та мінливості даних, що робить їх добре придатними для реальних застосувань, де виявлення підробок текстів має вирішальне значення для збереження довіри та цілісності в поширенні інформації.

У системах виявлення фейкових новин у режимі реального часу організовано безперервний процес для швидкої і точної оцінки достовірності новинних статей. Він починається з пошуку інформації, коли веб-пошукові системи розгортаються для постійного отримання новинних статей з різних онлайн-джерел, відповідаючи на запити користувачів або на заздалегідь визначені теми, що їх цікавлять. Ці пошукові роботи сканують Інтернет, щоб забезпечити систему найсвіжішою інформацією [46].

Після отримання статей у гру вступають NLP. Ці методи попередньо обробляють текстові дані, очищаючи їх від шуму та нерелевантної інформації, одночасно виділяючи релевантні ознаки, що вказують на автентичність статті. Для обробки тексту та вилучення значущої інформації використовуються такі методи, як токенізація, тегування частинами мови та розпізнавання іменованих об'єктів.

Після попередньої обробки до прогнозування залучаються моделі машинного навчання, навчені виявляти фейкові новини. Використовуючи витягнуті ознаки, ці моделі в режимі реального часу класифікують статті як

справжні, фейкові або підозрілі. Для підвищення точності та надійності прогнозування можуть застосовуватися ансамблеві методи або гібридні архітектури, які поєднують кілька моделей.

Інтеграція цих компонентів у цілісну систему дає змогу в режимі реального часу оцінювати достовірність новинних статей під час їхньої публікації в Інтернеті. Користувачі можуть робити запити до системи на певні теми або отримувати сповіщення про потенційно шахрайські новини, що дає їм змогу приймати обґрунтовані рішення та ефективно орієнтуватися в цифровому інформаційному ландшафті. Постійний моніторинг та оновлення гарантують, що система завжди реагує на нові тенденції та еволюцію тактик, які використовують зловмисники для розповсюдження дезінформації.

Для розгортання та підтримки системи виявлення фейкових новин у режимі реального часу необхідно дотримуватися кількох ключових стратегій.

По-перше, слід розробити зручний веб-додаток для полегшення взаємодії. Цей додаток дозволяє користувачам вводити новини або URL-адреси, ініціюючи процес класифікації. Потім система видає результат класифікації із зазначенням того, чи є стаття справжньою, фейковою або підозрілою. Крім того, програма може надавати пов'язані новини або додатковий контекст, щоб допомогти користувачам оцінити достовірність контенту [47].

По-друге, для безперешкодної інтеграції з іншими платформами або сервісами слід передбачити інтерфейси прикладного програмування (API). Ці API дають змогу розробникам вбудовувати функціональність системи виявлення фейкових новин у свої додатки або робочі процеси. Платформи соціальних мереж, агрегатори новин або системи управління контентом можуть інтегрувати ці API для автоматичного позначення потенційно оманливого контенту в режимі реального часу.

Безперервний моніторинг продуктивності системи має вирішальне значення для підтримки її ефективності та надійності. Це передбачає відстеження таких ключових показників, як точність класифікації, час відгуку та безперебійність роботи системи. Механізми моніторингу також повинні

виявляти будь-які відхилення або погіршення в роботі моделей, ініціюючи оновлення або перенавчання моделей за необхідності. Регулярне оновлення новими даними гарантує, що моделі залишатимуться актуальними та адаптованими до нових тенденцій і тактик, які застосовують зловмисники.

При розробці користувацького інтерфейсу (UI) та користувацького досвіду (UX) для системи виявлення фейкових новин у режимі реального часу вирішальне значення мають два ключові елементи: впровадження механізму зворотного зв'язку та візуалізація. Механізм зворотного зв'язку дозволяє користувачам повідомляти про помилкову класифікацію, надаючи цінну інформацію для покращення роботи системи. Візуалізація покращує розуміння та залучення користувачів завдяки візуальному представленню результатів виявлення та пов'язаних з ними новинних статей. Всі ці елементи разом сприяють створенню інтуїтивно зрозумілої та інформативної платформи для оцінки достовірності новинних статей в режимі реального часу.

Отже, виконавши всі етапи розробки та впровадження системи машинного навчання для виявлення фейків, матимемо надійний інструмент для роботи. Від збору даних та їх обробки до тренування моделей і розгортання системи – кожен етап важливий для забезпечення ефективної роботи системи. Додавання засобів зворотного зв'язку та візуалізації результатів дозволяє покращити взаємодію користувачів з системою та забезпечити її надійність в реальному часі.

3.3 Propaganda Detector як система машинного навчання для виявлення фейків у текстовому контенті

Propaganda Detector – система на основі штучного інтелекту, яка визначає техніки маніпуляції та пропаганди в тексті. Система здатна аналізувати тексти новин, статей та інших видів текстового контенту. Вона може виявляти до 18 різних технік пропаганди, що дозволяє детально аналізувати маніпулятивні

методи. Це допомагає аналітикам швидко і точно ідентифікувати пропагандистські елементи та розробити стратегії протидії.

Розглянемо інтерфейс системи, що зображений на Рис. 3.2.

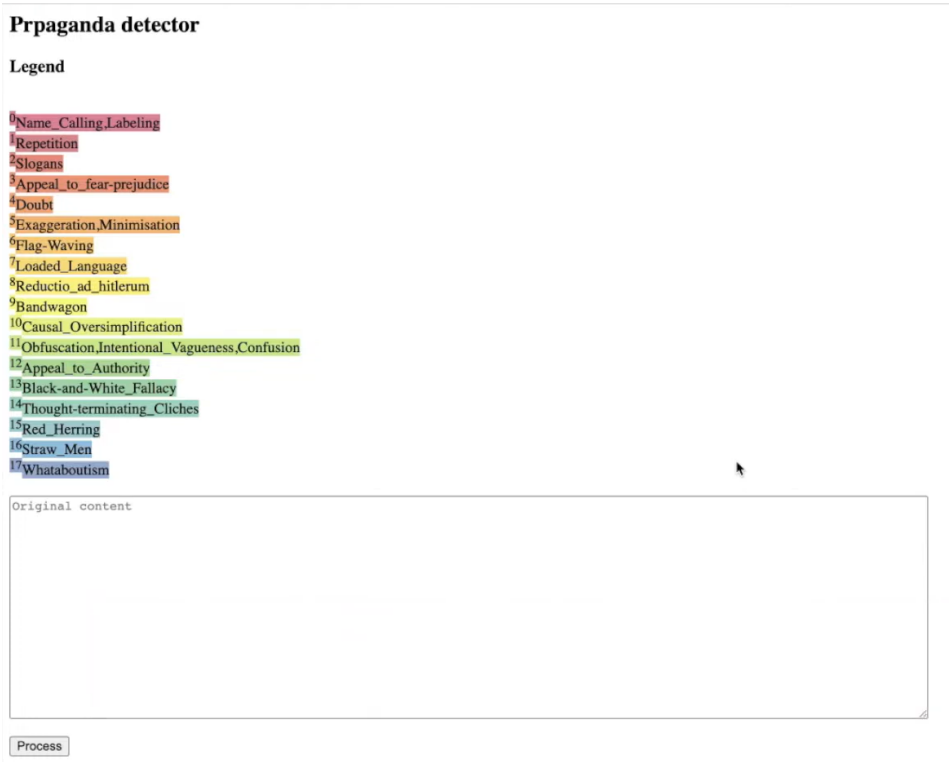

Рис. 3.2. Інтерфейс Propaganda Detector

На зображенні системи представлено кілька важливих елементів. У верхній частині знаходиться заголовок "Propaganda Detector" та легенда, яка пояснює різні техніки пропаганди, що система може визначити. Ці техніки представлені у вигляді кольорових міток.

"Name Calling" або "Labeling" спрямована на дискредитацію осіб чи груп через негативні ярлики, формуючи негативне ставлення без обґрунтованих доказів. "Repetition" полягає в багаторазовому повторенні однієї ідеї або гасла для зміцнення її у свідомості. "Slogans" використовують короткі, запам'ятовувані фрази для швидкого впливу на емоції. "Appeal to fear-prejudice" експлуатує страхи та упередження для маніпулювання думками та діями. "Doubt" вводить неясність щодо певних фактів або тверджень, щоб посіяти сумніви. "Exaggeration, Minimisation" перебільшує або зменшує значення подій чи фактів для зміни їхнього сприйняття. "Flag-Waving" асоціює ідеї з

патріотизмом чи національними символами для викликання патріотичних почуттів. "Loaded Language" використовує слова з сильним емоційним забарвленням для формування певних емоцій. "Reductio ad Hitlerum" порівнює опонентів або їхні аргументи з нацистами чи Гітлером для дискредитації. "Bandwagon" стимулює прийняття ідеї через те, що багато людей вже її підтримують. "Causal Oversimplification" пропонує надто прості причини складних проблем. "Obfuscation, Intentional Vagueness, Confusion" створює навмисну плутанину для приховування правди. "Appeal to Authority" посилається на авторитетні джерела для підтвердження тверджень, незалежно від їхньої релевантності. "Black-and-White Fallacy" пропонує лише два варіанти рішення, ігноруючи можливі альтернативи. "Thought-terminating Cliches" використовує прості фрази для припинення обговорення. "Red Herring" відволікає увагу на незначні деталі, відводячи від головного питання. "Straw Men" створює спотворене уявлення про аргументи опонента для їх легшого спростування. "Whataboutism" відволікає увагу від критики, піднімаючи інше питання, часто не пов'язане з темою. Ці техніки спрямовані на формування певних емоцій, думок та поведінки аудиторії, часто без належного обґрунтування чи логіки.

В центральній частині інтерфейсу розміщене текстове поле з підписом "Original content", куди користувач може вставити текст для аналізу. Під текстовим полем знаходиться кнопка "Process", яка запускає аналіз тексту на наявність технік пропаганди.

Підвантаживши текст, система розпізнає текст та аналізує його на задані техніки. Результат показано на Рис. 3.3.

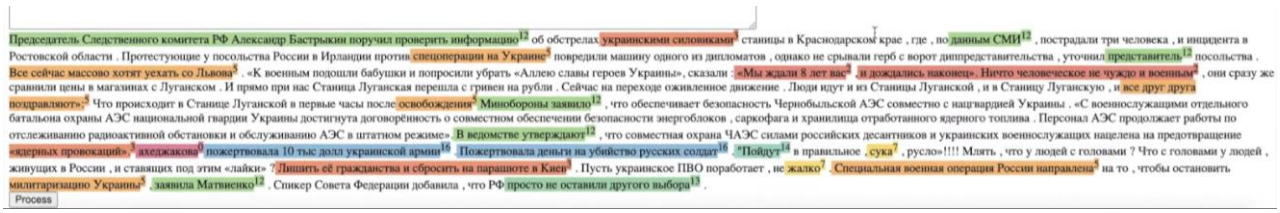

Рис. 3.3 Результат проведеного аналізу системою

В результаті завантажений текст було частково виділено різними кольорами та маркуваннями. Наприклад, система ідентифікувала техніку "Appeal to Authority" у фразі "по данным СМИ", де авторитетність джерела використовується для надання довіри до твердження. Фраза "Мы ждали 8 лет вас" була виділена як "Slogans" завдяки її емоційному та легко запам'ятовуваному характеру. "Exaggeration, Minimisation" видно у твердженні "все сейчас массово хотят уехать со Львова", де перебільшується масштаб явища. Інші техніки, такі як "Whataboutism" та "Bandwagon", також виявлені у тексті для демонстрації маніпулятивних практик.

Підсумовуючи, система "Propaganda Detector" ефективно ідентифікує різні техніки пропаганди у текстах за допомогою штучного інтелекту. Вона виділяє маніпулятивні стратегії, такі як звернення до авторитету, слогани, перебільшення, навантажена мова та відволікання уваги. Це дозволяє аналітикам глибше розуміти маніпуляції та розробляти стратегії протидії, забезпечуючи критичний аналіз інформаційного контенту для запобігання дезінформації.

Висновки до розділу 3

Розділ присвячений аналізу методів виявлення фейкових новин та створенню ефективної системи для їх ідентифікації. Цей розділ охоплює порівняння різних підходів та детальний опис процесу розробки системи виявлення фейків.

Порівняння та визначення найефективніших методів виявлення фейків є ключовим етапом у створенні системи машинного навчання. У цьому підпункті проводиться аналіз різних алгоритмів та технік, які використовуються для виявлення фейкових новин, таких як методи обробки природної мови (NLP), розумний аналіз даних та нейронні мережі. Порівняння включає оцінку точності, швидкості обробки та здатності адаптуватися до нових видів фейків. Дослідження показали, що поєднання кількох методів може дати кращі

результати, ніж використання одного окремого методу. Наприклад, поєднання методів NLP з нейронними мережами, такими як трансформери або LSTM, забезпечує високу точність та здатність враховувати контекст тексту. Також важливо враховувати такі фактори, як обчислювальні ресурси та складність реалізації, щоб вибраний метод був не лише ефективним, але й практичним для впровадження в реальних умовах.

Розробка системи виявлення фейків включає кілька етапів, починаючи з підготовки даних і закінчуючи тестуванням та впровадженням системи. На першому етапі здійснюється збір та попередня обробка даних, що включає очищення, нормалізацію та анотацію текстових корпусів, які використовуються для навчання моделі. Наступний етап полягає у виборі та налаштуванні алгоритмів машинного навчання. Для цього створюються та тренуються різні моделі, які порівнюються за показниками продуктивності, такими як точність, повнота та F-мера. Потім обирається найбільш ефективна модель або комбінація моделей.

Після цього відбувається етап інтеграції моделі в систему, яка включає розробку програмного забезпечення для автоматизованого аналізу текстового контенту. Це включає розробку інтерфейсів для вводу даних, модулів для обробки тексту, класифікації та зберігання результатів аналізу. Також важливим етапом є оптимізація системи для забезпечення її стабільної роботи та високої продуктивності в умовах реального часу.

Заключний етап розробки системи виявлення фейків полягає у її тестуванні та вдосконаленні. Система проходить через ряд тестів, включаючи перевірку на різних наборах даних для оцінки її здатності розпізнавати фейки в різних контекстах. На основі результатів тестування система може бути додатково налаштована для підвищення її точності та надійності. Після успішного тестування система впроваджується в експлуатацію, де продовжує навчатися та адаптуватися до нових викликів, пов'язаних з появою нових видів фейкових новин.

Тобто розділ демонструє комплексний підхід до створення ефективної системи для виявлення фейкових новин, починаючи від порівняння методів та вибору найбільш ефективних, до розробки, тестування та впровадження готової системи. Цей підхід забезпечує високу точність та надійність виявлення фейків у текстовому контенті, що є критично важливим у сучасному інформаційному середовищі.

ВИСНОВКИ

Розробка системи машинного навчання для виявлення фейків у текстовому контенті є важливим кроком у боротьбі з дезінформацією. Аналіз та порівняння різних методів виявлення фейків, таких як обробка природної мови, розумний аналіз даних та нейронні мережі, показали, що поєднання цих підходів може значно підвищити точність та ефективність системи. Процес розробки включає збір та підготовку даних, вибір і налаштування моделей, інтеграцію в програмне забезпечення та всебічне тестування. Успішна реалізація такої системи дозволяє автоматизувати процес ідентифікації фейкових новин, знижуючи їх поширення та мінімізуючи негативний вплив на суспільство.

Перший розділ теоретичні основи розуміння феномену фейкових новин. Визначення фейків як неправдивої інформації, що створена з метою маніпуляції, підкреслює важливість чіткого розмежування між правдивими та неправдивими даними. Включення в аналіз методів обробки природної мови (NLP) та роль машинного навчання акцентує на необхідності використання передових технологій для аналізу великих обсягів текстової інформації. Дослідження різних підходів до виявлення фейків, проведені провідними науковими інституціями, підтверджують значення міждисциплінарного підходу, що поєднує інформатику, соціологію та лінгвістику.

Другий розділ надає детальний аналіз сучасних технік для ідентифікації фейкових новин. Метод обробки природної мови (NLP) дозволяє здійснювати глибокий аналіз тексту, виявляючи лінгвістичні патерни, характерні для дезінформації. Розумний аналіз даних спрямований на ідентифікацію закономірностей у великих текстових масивах, що дозволяє прогнозувати поширення фейків та оцінювати їхній вплив на суспільство. Використання нейронних мереж, зокрема глибоких нейронних мереж, забезпечує високий рівень точності в класифікації текстів, дозволяючи ефективно розрізняти правдиву та неправдиву інформацію.

Третій розділ описує практичні аспекти створення системи для виявлення фейкових новин. Порівняння різних методів виявлення фейків підкреслює важливість вибору найбільш ефективних алгоритмів для конкретних завдань. Розробка системи включає етапи підготовки даних, вибору та налаштування моделей, інтеграції в програмне забезпечення та тестування. Впровадження такої системи в реальних умовах дозволяє автоматизувати процес виявлення фейків, підвищуючи ефективність боротьби з дезінформацією.

Кваліфікаційна робота підкреслює необхідність комплексного підходу до вирішення проблеми фейкових новин. Теоретичні основи, передові методи аналізу тексту та практичні аспекти розробки систем машинного навчання створюють основу для ефективної боротьби з дезінформацією. Інтеграція цих підходів забезпечує можливість не лише виявляти фейки з високою точністю, але й передбачати їх поширення та мінімізувати їхній негативний вплив на суспільство.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Collins 2017 Word of the Year Shortlist. *Collinsdictionary*. URL: https://blog.collinsdictionary.com/language-lovers/collins-2017-word-of-the-year-shortlist/

2. N. Belloir, W. Ouerdane, O. Pastor, É. Frugier, L.-A. de Barmon. A Conceptual Characterization of Fake News: A Positioning Paper. Research Challenges in Information Science. 2022. URL: https://link.springer.com/chapter/10.1007/978-3-031-05760-1_41

3. David Buckingham. Teaching Media in a 'Post-Truth' Age: Fake News, Media Bias and the Challenge for Media Literacy Education. *Cultura y Educación*. 2019. Vol. 31(2). URL: https://davidbuckingham.net/wp-content/uploads/2019/10/post-truth-buckingham.pdf

4. Fallis D., Mathiesen K. Fake news is counterfeit news. *Inquiry*. 2019. P. 1–20. URL: https://doi.org/10.1080/0020174x.2019.1688179

5. Michaelson E., Sterken R., Pepp J. What's New About Fake News?. *Journal of Ethics and Social Philosophy*. 2019. Vol. 16, no. 2. URL: https://doi.org/10.26556/jesp.v16i2.629

6. Tandoc E. C., Lim Z. W., Ling R. Defining "Fake News". *Digital Journalism*. 2017. Vol. 6, no. 2. P. 137–153. URL: https://doi.org/10.1080/21670811.2017.1360143

7. Rahmanian E. Fake news: a classification proposal and a future research agenda. *Spanish Journal of Marketing - ESIC*. 2022. URL: https://doi.org/10.1108/sjme-09-2021-0170

8. C. Wardle. Information disorder: Toward an interdisciplinary framework for research and policy making. 2017. URL: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

9. Yadav R. A Simplistic Overview of Machine Learning. *International Journal of Scientific Research in Science, Engineering and Technology*. 2021. P. 184–187. URL: https://doi.org/10.32628/ijsrset218475

10. B. Sharma. An Explanation of Machine Learning. *International Journal of Computer Science and Mobile Computing*. 2019. Vol.8 Issue.4. URL: https://ijcsmc.com/docs/papers/April2019/V8I4201918.pdf

11. Five machine learning types to know. *IBM*. URL: https://www.ibm.com/blog/machine-learning-types/

12. S. Mohamed, R. Ashraf, A. Ghanem, M. Sakr. Supervised Machine Learning Techniques: A Comparison. 2022. URL: https://www.researchgate.net/publication/363870735_Supervised_Machine_Learning_Techniques_A_Comparison

13. Comparing different supervised machine learning algorithms for disease prediction / S. Uddin et al. *BMC Medical Informatics and Decision Making*. 2019. Vol. 19, no. 1. URL: https://doi.org/10.1186/s12911-019-1004-8

14. Self-supervised learning: The dark matter of intelligence. *Meta*. 2021. URL: https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

15. C A Padmanabha Reddy Y., Viswanath P., Eswara Reddy B. Semi-supervised learning: a brief review. *International Journal of Engineering & Technology*. 2018. Vol. 7, no. 1.8. P. 81. URL: https://doi.org/10.14419/ijet.v7i1.8.9977

16. B. Dhanalaxm. Machine Learning and its Emergence in the Modern World and its Contribution to Artificial Intelligence. *2020 International Conference for Emerging Technology (INCET)*. 2020. URL: https://ksra.eu/wp-content/uploads/2020/11/10.1109@INCET49848.2020.9154058.pdf

17. Emerging technologies whitepaper series: Machine Learning. *Office of information and communication technology*. 2018. URL: https://unite.un.org/sites/unite.un.org/files/emerging-tech-series-machine-learning.pdf

18. T. R.A. Bakker. Objects in the Age of Virtual Reproduction: Aura and the Elusive Third Axis. 2018. URL: https://core.ac.uk/download/pdf/157730468.pdf

19. R. Dini. An Analysis of Walter Benjamin's The Work of Art in the Age of Mechanical Reproduction. 2017. URL: https://dokumen.pub/an-analysis-of-walter-benjamins-the-work-of-art-in-the-age-of-mechanical-reproduction-9781912304042-9781912284757-9781912284894.html

20. C. A. Gislam. The Effect of the Non-Human on the Generation of Narrative and Space in Digital Games. *Department of English Manchester Metropolitan University*. 2023. URL: https://e-space.mmu.ac.uk/634136/1/Charlotte_Gislam_PhD.pdf

21. E. Avilés. My/Mi lengua franca: Language, Manipulation, and Cultural Heritage in Chicana Art and Literature. 2014. URL: http://digitalrepository.unm.edu/span_etds/4

22. Manaris B. Natural Language Processing: A Human-Computer Interaction Perspective. *Advances in Computers*. 1998. P. 1–66. URL: https://doi.org/10.1016/s0065-2458(08)60665-8

23. Sawicki J., Ganzha M., Paprzycki M. The state of the art of Natural Language Processing - a systematic automated review of NLP literature using NLP techniques. *Data Intelligence*. 2023. P. 1–47. URL: https://doi.org/10.1162/dint_a_00213

24. Mah P. M., Skalna I., Muzam J. Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0. *Applied Sciences*. 2022. Vol. 12, no. 18. P. 9207. URL: https://doi.org/10.3390/app12189207

25. Perifanis N.-A., Kitsios F. Investigating the Influence of Artificial Intelligence on Business Value in the Digital Era of Strategy: A Literature Review. *Information*. 2023. Vol. 14, no. 2. P. 85. URL: https://doi.org/10.3390/info14020085

26. State of Art for Semantic Analysis of Natural Language Processing / D. Hussen Maulud et al. *Qubahan Academic Journal*. 2021. Vol. 1, no. 2. URL: https://doi.org/10.48161/qaj.v1n2a44

27. Chadha N., Gangwar R. C., Bedi R. Current Challenges and Application of Speech Recognition Process using Natural Language Processing: A Survey. *International Journal of Computer Applications*. 2015. Vol. 131, no. 11. P. 28–31. URL: https://doi.org/10.5120/ijca2015907471

28. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends / W. Khan et al. *Natural Language Processing Journal*. 2023. P. 100026. URL: https://doi.org/10.1016/j.nlp.2023.100026

29. Speech Recognition Using Deep Neural Networks: A Systematic Review / A. B. Nassif et al. *IEEE Access*. 2019. Vol. 7. P. 19143–19165. URL: https://doi.org/10.1109/access.2019.2896880

30. D. W. Otter, J. R. Medina, J. K. Kalita. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE transactions on neural networks and learning systems*. 2019. Vol. 20, № 10. URL: https://arxiv.org/pdf/1807.10854

31. Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges / S. O. Abioye et al. *Journal of Building Engineering*. 2021. Vol. 44. P. 103299. URL: https://doi.org/10.1016/j.jobe.2021.103299

32. Mahyoob M., Algaraady J., Alrahaili M. Linguistic-Based Detection of Fake News in Social Media. *International Journal of English Linguistics*. 2020. Vol. 11, no. 1. P. 99. URL: https://doi.org/10.5539/ijel.v11n1p99

33. P. K. Verma, P. Agrawal, I. Amorim, R. Prodan. WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE transactions on computational social systems*. 2021. Vol. 8, № 4. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9395133

34. Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges / N. R. de Oliveira et al. *Information*. 2021. Vol. 12, no. 1. P. 38. URL: https://doi.org/10.3390/info12010038

35. García-Díaz J. A., Beydoun G., Valencia-García R. Evaluating Transformers and Linguistic Features integration for Author Profiling tasks in

Spanish. *Data & Knowledge Engineering*. 2024. P. 102307. URL: https://doi.org/10.1016/j.datak.2024.102307

36. Linguistic Features and Bi-LSTM for Identification of Fake News / A. A. Ali et al. *Electronics*. 2023. Vol. 12, no. 13. P. 2942. URL: https://doi.org/10.3390/electronics12132942

37. J. Tompkins. Disinformation Detection: A review of linguistic feature selection and classification models in news veracity assessments. 2018. URL: https://arxiv.org/pdf/1910.12073

38. A Comprehensive survey of Fake news in Social Networks: Attributes, Features, and Detection Approaches. / M. R. Kondamudia et al. *Journal of King Saud University - Computer and Information Sciences*. 2023. P. 101571. URL: https://doi.org/10.1016/j.jksuci.2023.101571

39. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity / M. Alawida et al. *Information*. 2023. Vol. 14, no. 8. P. 462. URL: https://doi.org/10.3390/info14080462

40. LSTMCNN: A hybrid machine learning model to unmask fake news / D. G. Dev et al. *Heliyon*. 2024. Vol. 10, no. 3. P. e25244. URL: https://doi.org/10.1016/j.heliyon.2024.e25244

41. Nasir J. A., Khan O. S., Varlamis I. Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*. 2021. Vol. 1, no. 1. P. 100007. URL: https://doi.org/10.1016/j.jjimei.2020.100007

42. Mane S. A. M., Shinde A. StressNet: Hybrid model of LSTM and CNN for stress detection from electroencephalogram signal (EEG). *Results in Control and Optimization*. 2023. P. 100231. URL: https://doi.org/10.1016/j.rico.2023.100231

43. Sustainable Development of Information Dissemination: A Review of Current Fake News Detection Research and Practice / L. Yuan et al. *Systems*. 2023. Vol. 11, no. 9. P. 458. URL: https://doi.org/10.3390/systems11090458

44. A Comprehensive Review on Fake News Detection With Deep Learning / M. F. Mridha et al. *IEEE Access*. 2021. Vol. 9. P. 156151–156170. URL: https://doi.org/10.1109/access.2021.3129329

45. Taye M. M. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*. 2023. Vol. 12, no. 5. P. 91. URL: https://doi.org/10.3390/computers12050091

46. Fake news detection: a systematic literature review of machine learning algorithms and datasets / H. F. Villela et al. *Journal on Interactive Systems*. 2023. Vol. 14, no. 1. P. 47–58. URL: https://doi.org/10.5753/jis.2023.3020

47. Sentiment Analysis for Fake News Detection / M. A. Alonso et al. Electronics. 2021. Vol. 10, no. 11. P. 1348. URL: https://doi.org/10.3390/electronics10111348