МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ
КАФЕДРА УПРАВЛІННЯ КІБЕРБЕЗПЕКОЮ ТА ЗАХИСТОМ ІНФОРМАЦІЇ**

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «МОДЕЛЮВАННЯ ПРОЦЕСІВ ПОШИРЕННЯ ДЕЗІНФОРМАЦІЇ ТА МЕТОДІВ ЇЇ НЕЙТРАЛІЗАЦІЇ»

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

_____ Вікторія ОНІЩЕНКО
(підпис) *Ім'я, ПРІЗВИЩЕ здобувача*

Виконав(ла): здобувач(ка) вищої освіти гр. УБД-42

Вікторія ОНІЩЕНКО
Ім'я, ПРІЗВИЩЕ

Керівник: Тетяна БЕРЕСТЯНА
Ім'я, ПРІЗВИЩЕ

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

Київ 2025

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут кібербезпеки та захисту інформації

Кафедра управління кібербезпекою та захистом інформації

Ступінь вищої освіти бакалавр

Спеціальність 125 Кібербезпека

Освітня програма Управління інформаційною та кібернетичною безпекою

ЗАТВЕРДЖУЮ
Завідувач кафедри УКБЗІ
_____ Світлана ЛЕГОМІНОВА
“____ ”_____2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Оніщенко Вікторії Олексіївни
(прізвище, ім’я, по батькові здобувача)

1. Тема кваліфікаційної роботи “Моделювання процесів поширення дезінформації та методів її нейтралізації ”,
керівник кваліфікаційної роботи БЕРЕСТЯНА Тетяна
(ІПРІЗВИЩЕ, Ім’я,, науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологійвід “24“ лютого 2025 р. №56.

2. Строк подання кваліфікаційної роботи “12” травня 2025 р.
3. Вихідні дані до кваліфікаційної роботи: *цифрове середовище як простір поширення дезінформації; фейкові акаунти, боти та інформаційні кампанії; моделі поширення дезінформації;методи машинного навчання й штучного інтелекту.*
4. Перелік питань, які мають бути розроблені:
   4.1. Проаналізувати природу та особливості поширення дезінформації в соціальних мережах, класифікувати основні типи фейкових акаунтів і бот-мереж.
   4.2. Дослідити існуючі методи виявлення дезінформації, розглянути технології й інструменти, що застосовуються для аналізу даних у цифровому середовищі, оцінити їх ефективність.
   4.3. Розробити модель поширення дезінформації, що дозволяє врахувати ключові параметри інформаційних хвиль, запропонувати шляхи нейтралізації впливу дезінформації, зокрема шляхом виявлення фейкових джерел і мереж.
5. Перелік ілюстративного матеріалу: *презентація PowerPoint*
6. Дата видачі завдання “5“ березня 2025 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Етапи кваліфікаційної роботи	Термін виконання етапів роботи	Примітка
1.	Визначення об’єкту, предмету, мети та завдань дослідження.	18.03.2025	
2.	Збір та аналіз літератури.	29.03.2025	
3.	Аналіз особливостей управління інформаційною безпекою підприємства	08.04.2025	
4.	Дослідження основних характеристик технологій формування обізнаності й навчання персоналу.	15.04.2025	
5.	Вивчення інструментів та методів формування обізнаності й навчання персоналу з інформаційної безпеки	22.04.2025	
6.	Формулювання висновків за результатами проведеного дослідження.	29.04.2025	
7.	Оформлення роботи.	06.05.2025	
8.	Оформлення презентації.	09.05.2025	
9.	Отримання рецензії на роботу.	12.06.2025	
10.	Захист в ДЕК.	__.06.2025	

Здобувачка вищої освіти _____ Вікторія ОНІЩЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Керівник кваліфікаційної роботи _____ Тетяна БЕРЕСТЯНА
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

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

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

Направляється здобувачка Оніщенко В.О. до захисту кваліфікаційної роботи
(прізвище та ініціали)
за спеціальністю 125 Кібербезпека
(код, найменування спеціальності)
освітньої програми Управління інформаційною та кібернетичною безпекою
(назва)
на тему: “Моделювання процесів поширення дезінформації та методів її нейтралізації”
Кваліфікаційна робота і рецензія додаються.

Директор ННІКБЗІ _____ Євгенія ІВАНЧЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувачка ОНІЩЕНКО Вікторія у кваліфікаційній роботі комплексно дослідила проблему поширення дезінформації в цифровому середовищі. Проаналізовано механізми розповсюдження фейкової інформації, роль фейкових акаунтів і бот-мереж, а також підходи до моделювання дезінформаційних хвиль. Розглянуто сучасні технології виявлення та нейтралізації дезінформації із застосуванням методів аналізу даних і машинного навчання.

ОНІЩЕНКО Вікторія показала розуміння проблеми дослідження та бачення основних теоретичних і практичних напрямів її вирішення, довела володіння методами наукового дослідження, проявила себе як організована, відповідальна виконавиця. Результати дослідження апробовані на двох конференціях.

Все це дозволяє оцінити кваліфікаційну роботу здобувачки ОНІЩЕНКО Вікторії на оцінку “відмінно” та присвоїти їй кваліфікацію бакалавра з кібербезпеки за освітньою програмою Управління інформаційною та кібернетичною безпекою.

Керівник кваліфікаційної роботи _____ Тетяна БЕРЕСТЯНА
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

“____“_____2025 року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувачка Оніщенко В.О. допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедри управління
кібербезпекою та захистом
інформації _____ Світлана ЛЕГОМІНОВА
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну бакалаврську роботу

здобувачки вищої освіти ОНІЩЕНКО Вікторії
на тему: “Моделювання процесів поширення дезінформації та методів її нейтралізації”

Актуальність. Стрімке поширення цифрових технологій та глобальне зростання популярності соціальних мереж створюють сприятливі умови для масштабного розповсюдження дезінформації. Розробка ефективних методів моделювання процесів поширення дезінформації є актуальним завданням для сучасної науки, оскільки дозволяє виявляти закономірності її розповсюдження та розробляти стратегії її нейтралізації. Використання методів аналізу даних та штучного інтелекту відкриває нові можливості для виявлення інформаційних атак, оцінки рівня загрози та впровадження ефективних механізмів протидії.

З огляду на зазначене комплексне дослідження механізмів поширення дезінформації та розробка підходів до її обмеження є надзвичайно важливими як з наукової, так і з практичної точки зору.

Позитивні сторони.

1. У роботі досліджено механізми поширення дезінформації та запропоновано ефективні підходи до її нейтралізації.

2. Кваліфікаційна робота оформлена відповідно до вимог. Виклад матеріалу здійснено відповідно до плану, зроблено логічні висновки. Ключові положення роботи представлено у вигляді рисунків.

3. Авторка опрацювала значну джерельну базу: близько 49 публікацій, в тому числі англомовних.

4. За результатами дослідження запропоновано рекомендації рекомендації щодо вдосконалення інформаційного захисту від дезінформаційних впливів на основі сучасних аналітичних методів.

Недоліки.

Доцільно було б доповнити роботу глибшим аналізом практичних кейсів застосування моделей у реальних інформаційних кампаніях.

Однак, вищезгадані зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи.

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

Рецензент: _____ _____
підпис Ім’я, ПРІЗВИЩЕ

РЕФЕРАТ

Кваліфікаційна робота присвячена дослідженню процесів поширення дезінформації та методів її нейтралізації. Робота складається зі вступу, трьох розділів, що містять 9 рисунків, висновківі списку використаних джерел із 49 найменувань. Загальний обсяг роботи становить 85 аркушів, з яких 5 аркушів займають список використаних джерел.

Метою роботи є моделювання процесів розповсюдження дезінформації та створення підходів до її обмеження на основі методів аналізу даних і штучного інтелекту.

Об'єктом дослідження є процеси поширення дезінформації в інформаційному середовищі, зокрема в соціальних мережах, медіа та інших цифрових платформах.

Предмет дослідження – методи аналізу даних і штучного інтелекту для моделювання процесів поширення дезінформації та розробки підходів до її обмеження.

Методи дослідження. Для вирішення означеного вище наукового завдання в роботі використані методи аналізу та синтезу, порівняння, моделювання, та емпіричний аналіз поширення дезінформації

Як результат у роботі проаналізовано проблему дезінформації, змоделювано її поширення та запропонувано практичні рішення для її нейтралізації, що відповідає меті роботи та сучасним викликам у сфері кібербезпеки.

Галузь застосування. Запропоновані підходи можуть бути застосовані для протидії дезінформації в соціальних мережах та вдосконалення систем інформаційної безпеки в державному і приватному секторах.

КЛЮЧОВІ СЛОВА: ДЕЗІНФОРМАЦІЯ, СОЦІАЛЬНІ МЕРЕЖІ, ФЕЙКОВІ АКАУНТИ, ІНФОРМАЦІЙНИЙ ВПЛИВ, АНАЛІЗ ДАНИХ.

ABSTRACT

The qualification work is devoted to the study of the processes of disinformation dissemination and methods of its neutralization. The work consists of an introduction, three sections containing 9 figures, conclusions and a list of sources used from 49 items. The total volume of the work is 91 sheets, of which 4 sheets are occupied by a list of conventional abbreviations and a list of sources used.

The purpose of the study is to model the processes of disinformation dissemination and create approaches to its limitation based on data analysis methods and artificial intelligence.

The object of the study is the processes of disinformation dissemination in the information environment, in particular in social networks, media and other digital platforms.

The subject of the study is methods of data analysis and artificial intelligence for modeling the processes of disinformation dissemination and developing approaches to its limitation.

Research methods. To solve the above-mentioned scientific task, the work used methods of analysis and synthesis, comparison, modeling, and empirical analysis of the spread of disinformation. As a result, the work analyzed the problem of disinformation, modeled its spread, and proposed practical solutions for its neutralization, which corresponds to the purpose of the work and modern challenges in the field of cybersecurity.

Field of application. The proposed approaches can be applied to counter disinformation on social networks and to improve information security systems in both public and private sectors.

KEYWORDS: DISINFORMATION, SOCIAL MEDIA, FAKE ACCOUNTS, INFORMATION INFLUENCE, DATA ANALYSIS.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ І СКОРОЧЕНЬ

ШІ	Штучний інтелект
ML	Machine Learning (машинне навчання)
NLP	Natural Language Processing (обробка природної мови)
API	Application Programming Interface (інтерфейс прикладного програмування)
CNN	Згорткова нейронна мережа (англ. Convolutional Neural Network)
LSTM	Довга короткострокова пам’ять (англ. Long Short-Term Memory)
SVM	Метод опорних векторів (англ. Support Vector Machine)

ЗМІСТ

ВСТУП

Актуальність теми. Стрімке поширення цифрових технологій та глобальне зростання популярності соціальних мереж створюють сприятливі умови для масштабного розповсюдження дезінформації. Завдяки особливостям алгоритмів соціальних платформ, які сприяють вірусному поширенню контенту, недостовірна або маніпулятивна інформація може швидко охоплювати великі аудиторії, формуючи хибні уявлення та впливаючи на громадську думку.

Суттєву роль у цьому процесі відіграють як звичайні користувачі, що активно діляться неперевіреною інформацією, так і автоматизовані системи, зокрема боти та фейкові акаунти, які цілеспрямовано поширюють дезінформацію. Це особливо небезпечно в контексті політичних процесів, кризових ситуацій та інформаційних війн, коли маніпулятивний контент використовується для впливу на суспільні настрої та прийняття рішень.

Розробка ефективних методів моделювання процесів поширення дезінформації є актуальним завданням для сучасної науки, оскільки дозволяє виявляти закономірності її розповсюдження та розробляти стратегії її нейтралізації. Використання методів аналізу даних та штучного інтелекту відкриває нові можливості для виявлення інформаційних атак, оцінки рівня загрози та впровадження ефективних механізмів протидії.

Таким чином, комплексне дослідження механізмів поширення дезінформації та розробка підходів до її обмеження є надзвичайно важливими як з наукової, так і з практичної точки зору. Це сприятиме формуванню стійкого інформаційного середовища, що відповідатиме принципам свободи вираження поглядів, гарантованої статтею 10 Європейської конвенції про захист прав людини і основоположних свобод, та забезпеченню інформаційної безпеки суспільства.

Метою цієї роботи є моделювання процесів розповсюдження дезінформації та створення підходів до її обмеження на основі методів аналізу даних і штучного

інтелекту. Дослідження спрямоване на виявлення закономірностей розповсюдження фейкових новин, аналіз їхнього впливу на суспільну думку та інформаційну безпеку, а також створення ефективних стратегій нейтралізації дезінформаційних загроз.

Об'єкт дослідження Процеси поширення дезінформації в інформаційному середовищі, зокрема в соціальних мережах, медіа та інших цифрових платформах.

Предмет дослідження: Методи аналізу даних і штучного інтелекту для моделювання процесів поширення дезінформації та розробки підходів до її обмеження.

Для досягнення мети необхідно виконати такі **завдання**:

1. Проаналізувати сучасні підходи та механізми поширення дезінформації в цифровому середовищі.
2. Розробити модель поширення дезінформації з урахуванням впливу фейкових акаунтів і соціальних взаємодій.
3. Запропонувати методи нейтралізації дезінформації із застосуванням технологій аналізу даних і машинного навчання.

Методи дослідження. Для вирішення означеного наукового завдання в У роботі використано методи моделювання соціальних процесів, машинного навчання, аналізу соціальних мереж (SNA), обробки природної мови (NLP), методи статистичного аналізу, експертних оцінок та візуалізації даних.

Як результат, у роботі проведено аналіз наявних моделей поширення інформації в соціальних мережах, досліджено застосування технологій штучного інтелекту для виявлення дезінформаційного контенту, розроблено і протестовано модель поширення дезінформації з імітацією дій фейкових акаунтів, а також запропоновано практичні методи обмеження їхнього впливу.

Наукова новизна полягає у побудові адаптивної моделі поширення дезінформації з урахуванням індивідуальних параметрів користувачів і типів

дезінформаційних впливів, що дозволяє підвищити точність прогнозування динаміки інформаційних атак.

Практичне значення одержаних результатів. полягає у розробці моделей і рекомендацій, що можуть бути використані для побудови систем раннього виявлення та протидії дезінформаційним кампаніям у соціальних мережах, зокрема під час кризових ситуацій, гібридних загроз та інформаційних війн.

РОЗДІЛ 1. ТЕОРЕТИЧНІ АСПЕКТИ ПОШИРЕННЯ ДЕЗІНФОРМАЦІЇ

1.1. Поняття, класифікація та види дезінформації

Спеціальний доповідач ООН із заохочення і захисту права на свободу думок і їхнє вільне вираження Ірен Хан, у своїй доповіді від 13 квітня 2021 року зазначила: «Дезінформація – це не нове явище, але новим є те, що цифрові технології дозволили різним суб'єктам створювати, поширювати та посилювати неправдиву, чи підроблену інформацію в політичних, ідеологічних або комерційних цілях в небачених до цього масштабах та з такою швидкістю і межах доступності, про які раніше не було відомо» [1, с.33].

Не дивлячись на те, що в наші часи за допомогою інтернету набагато легше поширювати будь-яку інформацію і вона може впливати на події і людей швидше, ніж до появи сучасних ЗМІ, в нашій історії існує чимало випадків, коли дезінформація призводила до фатальних наслідків. Одним з таких випадків був Фальшивий "Протокол сіонських мудреців", який використовували для розпалювання антисемітизму та виправдання переслідувань. 1903 року уривки «*Протоколів сіонських мудреців*» були послідовно опубліковані в російській газеті «*Знамя*» (Прапор). У 24 розділах, або ж нібито протоколах зустрічей єврейських лідерів, «описано таємні плани» євреїв правити світом, керуючи економікою, контролюючи засоби масової інформації та розпалюючи релігійні конфлікти. 1921 року лондонське видання Times представило переконливі докази того, що «*Протоколи*» виявилися «незграбним плагіатом». Times підтвердила, що «*Протоколи*» були значною мірою скопійовані з французької політичної сатири, в якій євреї ніколи не згадувалися, – «*Діалогу в пеклі між Макіавеллі та Монтеск'є*» (1864) Моріса Жолі. Інші дослідження показали, що одна глава роману «*Біарріц*» прусського письменника Германа Гедше (1868) також стала «натхненням» для «*Протоколів*» [2].

Прототипні приклади дезінформації включають оманливу рекламу (у бізнесі та політиці), урядову пропаганду, підроблені фотографії, підроблені документи, підроблені карти, інтернет-шахрайство, підроблені веб-сайти та маніпульовані записи у Вікіпедії [3].

В контексті виявлення навмисно сфабрикованого контенту частіше за все можна зустріти такі ключові слова та термінологію: помилкова дезінформація (в українській мові я не можу знайти переклад у вигляді одного слова, який був би відповідником терміну «misinformation» в англійській мові), дезінформація, фейкові новини, обман, чутки та теорія змови.

Отже, переходжу до визначення цих термінів у спосіб, який використовували більшість дослідників.

Помилкова дезінформація («misinformation») [4] описується в літературі як «неправдива, помилкова або оманлива» інформація [5], яку часто вважають «чесною помилкою».

З іншого боку, дезінформація – це неправдива інформація, навмисно поширена з наміром ввести в оману та/або обдурити [6]. Фейкові новини визначено як «новинні статті, які є навмисно та підтверджено неправдивими та можуть ввести читачів в оману» [7], і більшість дослідників дотримуються цього визначення. Отже, фейкові новини є прикладом дезінформації.

Багато дослідників також використовують термін «містифікація» (походить від hocus, що означає обдурити), посилатися на навмисно неправдиву інформацію [8]. Містифікації — це повідомлення, створені з наміром поширити серед великої кількості людей, щоб «переконати або маніпулювати іншими людьми зробити або запобігти заздалегідь встановленим діям, переважно за допомогою погроз або обману».

Також до дезінформації відносяться «чутки». Відповідно до [9], «чутки» — це «публічні повідомлення, які переповнені приватними гіпотезами про те, як влаштований світ» [10].

Якщо брати конкретно дезінформацію, то вона також поділяється на певні види. При пошуку інформації щодо її видів я натрапляла на різні класифікації, бо поділити можна за різними ознаками: за метою, способом поширення, ступенем викривлення правди, джерелом поширення і т.д. Проте, найчастіше її поділяють на такі види:

1**.** Сатира чи пародія. Без шкідливих намірів, але потенційно оманливі.

2. Помилкове з'єднання. Коли заголовки, візуальні елементи чи підписи не відповідають змісту.

3. Введення в оману. Використання оманливої інформацію, щоб висвітлити проблему чи особу.

4. Помилковий контекст. Коли поширюється справжній вміст разом із неправдивою контекстною інформацією.

5. Самозванець. Коли видають себе за справжні джерела.

6. Маніпуляційний вміст. Коли справжньою інформацією або зображеннями маніпулюють з метою введення в оману.

7. Сфабрикований вміст. Новий контент на 100% неправдивий і має на меті обдурити та нашкодити [11].

Так як у наш час дезінформація переважно поширюється через соціальні мережі, то було б доречним вказати наскільки взагалі соцмережі стали популярними за останній рік і наскільки збільшилась кількість користувачів.

Згідно з даними сайту Data Reportal [12] детальний аналіз, здійснений командою Kepios, демонструє, що на початок квітня 2025 року в усьому світі було 5,31 мільярда користувачів соціальних мереж, що становить 64,7 відсотка всього населення світу.

Кількість користувачів соціальних медіа також продовжувала збільшуватися протягом останніх 12 місяців: 241 мільйон нових користувачів долучилися до соціальних мереж минулого року.

Це дорівнює річному зростанню на 4,7 відсотка за середньої швидкості 7,6 нових користувачів щосекунди.

Свіжі дані показують, що 94,2 відсотка користувачів Інтернету в світі щомісяця користуються соціальними мережами.

На рис.1 зображено соціальні мережі, які найчастіше використовуються у світі.

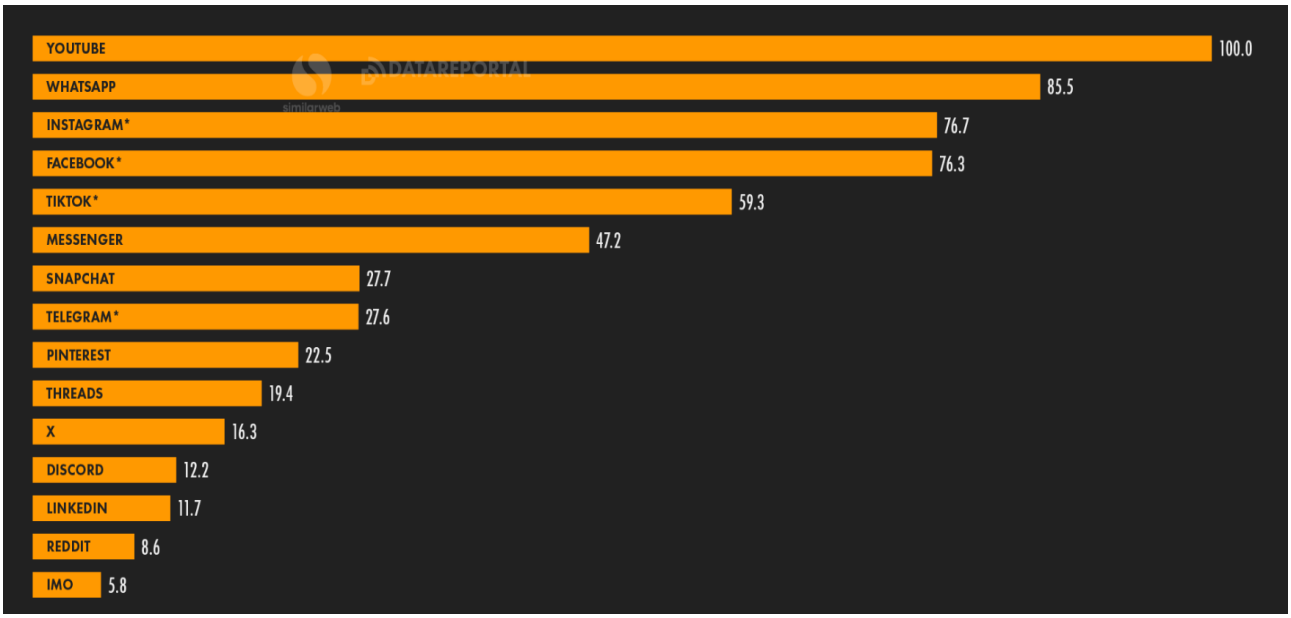

Рис. 1. Програми соціальних мереж: індекс активного користувача [12]

Оскільки люди проводять більше часу на цих сайтах, щоб отримати останні новини та інформацію, не можна недооцінювати їхню важливість у поширенні фейкових новин.

Соціальні мережі є одним з трьох основних елементів у поширенні дезінформації, згідно з [13].

Як показано на Рис. 2, загалом, фейкова новина складається з платформи для її поширення, тобто соціальних мереж, плюс інструменти та сервіси для маніпулювання та поширення і третій елемент це мотивація.

Рис. 2. Трикутник фейкових новин [13]

1.2 Основні механізми поширення фейкових новин та інформаційних маніпуляцій

Існує багато механізмів поширення дезінформації, далі розглянуто, та наведено приклади з реального життя, які пов’язані з війною в Україні [14]:

1. «Заговорювання» – цей метод використовується пропагандистами, коли їм потрібно применшити актуальність події та спровокувати негативну реакцію. Мета «заговорювання» - переситити аудиторію дезінформацією про певні «гарячі» теми і не дати їй більше ними цікавитися.

2. Ефект «первинності» – успіх пропаганди полягає в тому, що дезінформація досягає аудиторії раніше, ніж правда. Коли люди отримують суперечливу інформацію, яку неможливо негайно перевірити, вони схильні надавати перевагу тій, яка з'являється першою.

3. «Буденної розповіді» – щоб «ознайомити» глядачів з дезінформацією про насильство, вбивства, теракти та обстріли, російська пропаганда використовує

«буденну розповідь». «Благородні» телеведучі на каналах “Ростелекома” регулярно повідомляють про найтяжчі злочини, скоєні російськими військовими, зі спокійним виразом обличчя і рівним голосом. Через деякий час під впливом такої «обробки» росіяни припиняють реагувати на найжахливіші злочини та масові убивства, вчинені в Україні. Ефект полягає в тому, що вони «звикають» до насилля.

4. «Удар на випередження» – це випереджувальна інформаційна операція. Її мета - спровокувати контратаку «противника» і використати її з більшою вигодою для себе. Цей прийом використовується російськими пропагандистами для прискорення ескалації конфлікту і провокацій шляхом поширення дезінформації, тим самим «гасячи» правдиву інформацію.

5. «Хибної аналогії» – виникає при порівнянні властивостей і відносин, які не мають під собою ніякої основи в реальності. Люди схильні проводити аналогії, використовуючи логічні причинно-наслідкові зв'язки. Заміна цих зв'язків пропагандою призводить до «хибних аналогій», коли ймовірність висновку дорівнює нулю.

6. «Констатація факту» – це подача бажаної інформації в ЗМІ як беззаперечної. Часто цей механізм використовується у вигляді новин або соціологічних досліджень. Для надання більшої достовірності матеріалам, що використовують механізм «констатації факту», пропаганда широко використовує коментарі «лідерів думок» - популярних журналістів, відомих політологів тощо.

7. «Обхід з флангу» – це передача великих обсягів достовірної інформації, про яку слухачі та читачі знають заздалегідь і достовірність якої можна легко перевірити. Пропагандистські повідомлення фактично «упаковані» в такі дані. Мета використання «обходу з флангу» - привернути увагу аудиторії до знайомої інформації, щоб гарантувати, що вона отримає саме ту інформацію, яка їй потрібна. При «франкуванні» часто використовують назви вулиць і номери будинків. Пропагандистське повідомлення «непомітно» додається до багатьох достовірних деталей.

8. «Створення проблеми» – означає цілеспрямований підбір інформації з метою підвищення важливості однієї події та зниження її значущості по відношенню до інших подій.

9. «Керованих коментарів» – це повідомлення, в якому думки людини спрямовуються в потрібному напрямку. Як це працює: повідомлення супроводжується інтерпретацією коментатора, який пропонує читачеві кілька варіантів пояснень. Коментарі пишуться за принципом «двостороннього повідомлення» і містять аргументи як «за», так і «проти», залежно від вибору читача. Щоб зробити коментар правдоподібним, позитивні та негативні елементи чергуються, а факти підбираються таким чином, щоб посилити або послабити твердження чи настрої. Висновки не є частиною коментаря, оскільки вони робляться читачем під впливом коментаря.

10. «Ефект ореолу» – загалом суть психологічного «ефекту ореолу» можна виразити двома словами: «прославляння “величі” одного солдата в окупаційній армії» Прирівнюючи «велич» одного солдата до «величі» всього підрозділу, ефект «величі» підрозділу поширюється на всю армію загарбника: «героїчний воїн - ефективна армія».

11. «Відволікання уваги» – це перенесення зосередження з одного предмета на інший. Відбувається через поширення даних, які є стороннім подразником для особи, яка в цей момент зайнята важливою справою.

12. «Переконання» — це метод організованого впливу на психологічний стан особи, через звернення до її критичного мислення, з ідеєю, у яку людина непохитно вірить.

13. Маніпулювання опитуваннями громадської думки, або принцип «спіралі мовчання» – це підбір коментарів, що має переконати громадян у підтримці більшістю суспільства тієї чи іншої точки зору або позиції. Німецька політологиня Елізабет Ноель-Нойманн запропонувала визначення «спіралі мовчання», щоб пояснити, як люди вважають за краще промовчати, коли відчувають, що їхні

погляди в меншості, непопулярні, або це просто окрема думка, щоб не потрапити в соціально-психологічну ізоляцію.

14. «Зараження» або «стадний» інстинкт» – один зі способів впливу групи на психіку. Цей метод об'єднання групової діяльності виникає, коли людей дуже багато. Одним з проявів є стихійність – «всі побігли, і я теж». Під час психічного зараження емоції передаються від людини до людини несвідомо.

15. «Дублювання акаунтів» – це створення облікових записів, які копіюють вміст справжніх акаунтів користувачів та використовуються для шахрайства, поширення дезінформації та іншого. Між назвою справжнього та клонованого облікового запису іноді різниця лише в одному символі, що одразу помітно. Частіше за все, зловмисники клонують акаунти ключових державних діячів, блогерів, публічних осіб, які мають значний вплив на громадську думку.

16. «Підміна понять» – це психологічний прийом, заснований на логічній хибності. Він полягає у тому, щоб представляти аудиторії будь-який об’єкт чи явище як те, чим він не є. З часом цей підмінений вислів закріплюється та починає функціонувати у суспільстві як єдино правильний [14].

1.3. Вплив дезінформації на суспільну думку та інформаційну безпеку

Дезінформація може впливати на різні сфери нашого життя, на політику, здоров’я, бізнес і т.д., при цьому вона несе великі збитки як репутаційні так і фінансові, може керувати людьми чи навіть цілими державами. Я наведу деякі приклади з реального життя, пов’язані з цією проблемою.

Неправдива інформація, пов’язана зі здоров’ям, на мою думку, є найбільш небезпечною, вона може ввести громадськість в оману та зірвати програми громадської охорони здоров’я. Згідно зі статтею в журналі, опублікованому на Oxford Academic дезінформація про здоров’я, що охоплює низку тем, включаючи вакцини, інфекційні захворювання, харчування, зміну клімату, рак і куріння,

широко поширена на основних платформах соціальних мереж. Ще до пандемії COVID-19 відбулося посилення руху «anti-vaxx», головним чином завдяки поширенню дезінформації, головним чином у соціальних мережах. Дезінформація щодо вакцин призвела до зниження рівня вакцинації та спалахів захворювання, включно з кором, у регіонах, де було досягнуто ліквідації. Тоді пандемія вивела вакцинацію на перший план громадського дискурсу. Глобальні мережі дезінформації про здоров'я опублікували дезінформацію про COVID-19 і вакцини, що зібрало 3,8 мільярда переглядів у Facebook за 12 місяців. Дезінформація та теорії змови сприяли ваганням щодо вакцини проти COVID-19 і викликали посилення страху та занепокоєння протягом усієї пандемії.

Ключові галузі, чиї системи, практики та продукти визнані CDoH, також сприяють поширенню дезінформації в соціальних мережах. Тютюнова промисловість, яка історично вводила громадськість в оману щодо шкоди тютюнових виробів, продає нові нікотинові продукти (наприклад, електронні сигарети або «пари» та бездимний тютюн) через впливових людей у соціальних мережах, деякі з яких стверджують, що продукти нешкідливі. Тютюнова та інші шкідливі галузі, такі як індустрія засмаги, дедалі частіше представляють себе науковими авторитетами, використовуючи тактику дезінформації, зокрема підкреслюючи сфери наукової невизначеності та проводячи чи замовляючи дослідження, які служать їхнім інтересам (Reed et al. 2021). Експериментальне дослідження показало, що молоді дорослі, які бачили оманливий контент у соціальних мережах про вейпи, виявляли більш прихильне ставлення до вейпів порівняно з контрольною групою, що свідчить про те, що оманливий маркетинг може бути впливовою та тактикою, яку можуть використовувати шкідливі індустрії.

Описані вище обставини сприяють розмиванню довіри до сфери охорони здоров'я та науки. Дезінформація про здоров'я часто викликає розбіжності, і, як відомо, особи, які впливають на здоров'я, підривають авторитет органів охорони

здоров’я та науки. Така дезінформація, пов’язана зі здоров’ям, сприяє поляризації та недовірі громадськості до надійних експертів із охорони здоров’я та організацій. Медичні працівники та експерти з конфліктом інтересів, реальним чи уявним, також сприяють підриву довіри до громадської охорони здоров’я. Натомість багато людей довіряють альтернативним джерелам, які поширюють дезінформацію та дискредитують законних експертів у сфері охорони здоров’я, створюючи парадокс: недовіра, що підживлюється дезінформацією, сприяє її подальшому поширенню. Це проблематично, оскільки недовіра впливає на дотримання громадськістю рекомендацій щодо здоров’я, заснованих на фактичних даних, що може вплинути на здоров’я окремих людей і населення. Після того, як довіру було зруйновано, відновити її складно та довго. Крім того, поширеність дезінформації та підрив довіри до громадської охорони здоров’я також можуть служити інтересам шкідливих галузей, підриваючи та відволікаючи від обговорення регулювання та інших заходів [15].

Також, окрім дезінформації про здоров’я дуже небезпечною є неправдива інформація, пов’язана з війною, оскільки вона не тільки деморалізує людей, сіє паніку і хаос, а і може підірвати міжнародну підтримку. Так, напад ХАМАС на Ізраїль 7 жовтня 2023 року призвів до розповсюдження серії фейкових новин, які в основному стосуються того, що відбувається на місці. Повідомлення про обезголовлення хлопчика та викрадення десятків ізраїльських дівчат, які бійці ХАМАС використовували як сексуальних рабинь, уже підтвердили свою неправдивість. Ці фейкові новини, здається, надходять для переважної більшості з індійських акаунтів. Кореляція може полягати в тому, що, особливо після піднесення прем’єр-міністра Нарендри Моді та його партії Бхаратія Джаната (BJP), в Індії стався спалах ісламофобії. Наприклад, сім із десяти облікових записів X (колишній Twitter), які поширювали дезінформацію про ізраїльських секс-рабинь, були простежені до індійських профілів. Тільки ці сім твітів мали 3 мільйони показів на X (раніше Twitter). Відео було не тільки вересневим, але й простою

шкільною екскурсією в Єрусалимі. 17 жовтня 2023 року в лікарні Аль-Ахлі в Газі стався катастрофічний вибух, у результаті якого загинули понад 500 мирних жителів. Відразу палестинська влада звинуватила в нападі Ізраїль. Громадська думка в арабському світі стала на бік палестинців, засуджуючи Ізраїль у вчиненні військових злочинів. Однак ізраїльський уряд рішуче заперечував будь-яку причетність до вибуху, стверджуючи, що з невідомих причин ХАМАС завдав удару по його власній лікарні. Згідно з аналізом кадрів, проведеним ЦАХАЛом, кратер, утворений вибухом, був би занадто малим, щоб бути спричиненим ізраїльською зброєю. Крім того, ЦАХАЛ оприлюднив розмову, ймовірно, між двома бойовиками ХАМАС, у якій вони визнали, що завдали удару по лікарні Аль-Ахлі.

Автократичні держави використовують властиві слабкості демократичних систем за допомогою стратегій втручання через дезінформацію та кібератаки. Кінцева ціль – послаблення демократичних інституцій, підриваючи їхню внутрішню єдність, та маніпулюючи громадською думкою. Суперечки спрямовані на поляризацію та запрошення громадян обрати сторону. Цей тип операцій широко застосовує соціальні медіа, алгоритми котрих свідомо надають перевагу найбільш розбіжному та поляризуючому контенту [16].

Статтю про те, як брехня впливає на бізнес публікували у Financial Times [17], де писали, що уряди давно скаржаться на дезінформацію, яка підриває вибори, розпалює заворушення та поглиблює суспільний розкол. Підприємства також зараз під прицілом. «За останні кілька років дезінформація зростала, і тепер вона поширюється, як спориш», — сказав Джуліан Пейн, глобальний голова відділу криз та ризиків консалтингової компанії Edelman.

Для великих компаній управління наслідками внутрішніх помилок — від невдалих реакцій на політичні кризи до неправомірних дій керівництва — завжди було складним завданням. Але захист репутації стає лише складнішим.

Вже досить складно було, коли корпоративні кризи були більше пов'язані з реальними подіями чи рішеннями. Тепер брехня та дезінформація можуть збити

компанії з пантелику, оскільки онлайн-фейки можуть трансформуватися та розмножуватися швидше, ніж будь-коли раніше.

Arla Foods, власник найбільшого у Великій Британії молочного кооперативу, отримав гіркий досвід. Нещодавно оголосивши про випробування кормової добавки, спрямованої на зменшення викидів метану у молочних корів, деякі клієнти наполягали на бойкоті молочних продуктів. У соціальних мережах розгорілася буря, пов'язана з необґрунтованими та дивними твердженнями про те, що добавка була частиною змови з метою зменшення населення світу шляхом створення проблем з фертильністю, а також теоріями змови, що пов'язують добавку з Біллом Гейтсом. Британські регулятори схвалили добавку Bovaer як безпечну, тоді як її виробник звинуватив у цьому шаленстві «неправду та дезінформацію» [17].

Як наслідок, бойкот споживачів призвів до зниження продажів, відповідно це спричинило фінансові та репутаційні втрати.

1.4. Роль соціальних мереж, ботів та фейкових акаунтів у розповсюдженні дезінформації

Через те, що технології та соціальні медіа безперервно та швидко розвиваються, платформи можуть вносити, здавалося б, миттєві зміни в свій алгоритм або інтерфейс, це матиме значні наслідки для функціональності платформи для мільйонів користувачів. Останніми роками стався сплеск розвитку штучного інтелекту (ШІ), більша частина якого фінансується технологічними гігантами, зокрема Meta та Google. За прогнозами, до 2030 року штучний інтелект досягне рівня людського інтелекту і значно перевищить його в наступні роки, а платформи соціальних мереж, зокрема Instagram, LinkedIn і Snapchat, вбудували інструменти ШІ у свої платформи. Ці швидкі та безпрецедентні технологічні досягнення ускладнюють розробку та впровадження своєчасної та ефективної політики [18].

Для початку потрібно розібратись що таке «бот». Бот – це повністю або частково автоматизована система, яка виконує поставлені завдання. Відповідно до їхнього призначення розрізняють багато різновидів ботів, серед яких найпопулярнішими є:

- веб-сканери, що періодично переглядають мережу для індексації сторінок;
- чат-боти, використовують штучний інтелект та імітують персональну розмову, відомими прикладами є ChatGPT та Gemini, також боти, які автоматизують обслуговування клієнтів на веб-сайтах;
- соціальні боти, облікові записи призначені для одно- або багатостороннього спілкування та наслідування поведінки людини в соціальних мережах. Соціальні боти це керовані програмним забезпеченням облікові записи в соціальних мережах, які імітують людей зі зловмисними намірами [19].

Боти можуть автоматично створювати та поширювати дезінформацію та пропаганду з метою впливу на громадську думку, можуть автоматично розміщувати коментарі або повідомлення, тим самим перешкоджаючі спілкуванню користувачів та завдяки провокаційним висловлюванням розпалювати ворожнечу у соцмережах, можуть виманювати конфіденційну інформацію у користувачів. Ці форми діяльності ботів можуть мати серйозні наслідки для користувачів соціальних мереж та суспільства в цілому, тому важливо бути обережними у соцмережах та піддавати інформацію критичному мисленню. Багато підроблених облікових записів створюються з метою впливу на користувачів для просування своїх рекламних кампаній чи підтримки певних політичних партій [19].

Ще більш цікавий та небезпечний вид поширення дезінформації це використання технології deepfake (поєднання понять «deep learning» («глибоке навчання») та «fake»(«підробка»)). Як писали BBC, діпфейки – це відео, зображення чи аудіокліпи, створені за допомогою штучного інтелекту, щоб виглядати реалістично. Їх можна використовувати для розваги або навіть для наукових

досліджень, але іноді їх використовують для того, щоб видавати себе за інших людей, таких як політики чи світові лідери, щоб навмисно ввести людей в оману.

За допомогою штучного інтелекту діпфейки можуть імітувати голос та риси обличчя людини. Технологія використовує аудіозапис чийогось голосу, щоб змусити його говорити те, чого людина, можливо, ніколи б не сказала. Вона може імітувати рухи обличчя людини з відео з нею або навіть просто зображення її обличчя. Багато діпфейкових відео чи зображень можуть виглядати дуже дивно, тому їх легко розпізнати як імітації. Однак іноді вони можуть виглядати досить реалістично, і технології, які люди використовують для їх створення, постійно вдосконалюються [19].

Застосування таких прийомів робить телебачення значно могутнішим засобом для інформаційного впливу на споживачів інформації. Для підтвердження цієї тези можна використати свідченнями журналістів, які стикалися з наслідками використання сучасних технологій, за допомогою яких аудіо чи відеокліпи здатні діяти як дуже потужні «апробації інформаційного впливу». Один з прикладів таких технології це VoCo, який генерує голос людини і можна в реальному часі, під час перегляду відео з людиною, змінювати її слова. Для цього програма аналізує 20 або навіть 40 хвилин, якщо ви хочете найкращих результатів, вашої розмови, і вона розпізнає всю фонетику вашої мови, всі звуки, які ви вимовляєте. Знаходить кожен маленький блок звуку та мовлення в записах. Розбиває їх усі на частини. А потім, коли ви вводите щось, він об'єднує їх у нове слово [20].

Зловмисники, включаючи екстремістів та ворожі іноземні держави, використовують мережі фейкових акаунтів для поширення та посилення дезінформації, провокаційного та роз'єднуючого контенту. Ці фейкові акаунти здатні обманювати кінцевих користувачів неавтентичними публікаціями, коментарями та «лайками». Вони також використовуються для обману алгоритмів рекомендацій платформ, щоб дезінформація ширше поширювалася в стрічках новин користувачів. Дезінформація, що поширюється через фейкові акаунти,

підвищує рівень загрози для кандидатів та депутатів, підвищуючи напруженість навколо політичних дебатів та розпалюючи ворожість до політиків. Вона також може загрожувати цілісності виборчих процесів, спотворюючи дебати або підриваючи довіру до виборчих процесів. Напередодні останніх загальних виборів у Великій Британії, Об'єднаний комітет з питань стратегії національної безпеки наголосив на загрозах для виборів, які створює використання соціальних мереж ворожими іноземними державами для «поширення дезінформації в Інтернеті про публічних осіб, щоб підживлювати теорії змови та підривати довіру», а також для «сіяння розколу та сприяння хаосу там, де вже існують внутрішні розбіжності щодо суперечливих або політизованих тем».

У липні 2024 року Global Witness повідомила, що «45 облікових записів, які виглядають як боти на X, згенерували понад 4 мільярди переглядів, одночасно посилюючи расистські та сексуалізовані образи, теорії змови та дезінформацію щодо клімату». Ці облікові записи, орієнтовані на Велику Британію, спочатку були виявлені Global Witness через їхню спрямованість на загальні вибори у Великій Британії, але дослідники зазначили, що «акаунти, які активно публікували дописи під час загальних виборів у Великій Британії, швидко реагують на нові теми, що виникають, посилюючи суперечливий контент».

У грудні 2024 року президентські вибори в Румунії були призупинені після того, як було виявлено, що дезінформаційна кампанія мала серйозний, спотворюючий вплив. Фейкові акаунти були основною рисою цієї дезінформаційної операції – румунська розвідка виявила тисячі фальшивих акаунтів TikTok, які, здавалося, перебували під контролем Росії та підтримували ультраправого, проросійського кандидата. Європейська комісія розпочала розслідування DSA щодо очевидної нездатності TikTok вжити достатніх заходів для запобігання «неавтентичним маніпуляціям» його рекомендаційними системами фальшивими акаунтами.

Як і у випадку з анонімними акаунтами, Ofcom визнає цю проблему. У профілях ризиків він зазначає, що платформи, які дозволяють створювати анонімні акаунти, «ймовірно, мають підвищений ризик шкоди, пов'язаної зі злочинами, пов'язаними з іноземним втручанням» (а також іншими серйозними злочинами, включаючи домагання/переслідування/погрози/зловживання та шахрайство) [21].

Висновки до першого розділу 1

У першому розділі я здійснила теоретичне обґрунтування поняття дезінформації як явища, що має складну природу та суттєвий вплив на інформаційний простір. Встановила, що дезінформація є цілеспрямованим поширенням неправдивої або викривленої інформації з метою впливу на поведінку, свідомість або рішення окремих осіб, груп чи суспільства загалом.

Було проаналізовано основні класифікації та види дезінформації, що дозволяє краще розуміти її природу та специфіку використання в різних сферах. Окрему увагу приділила механізмам поширення фейкових новин та інформаційних маніпуляцій, які базуються на використанні психологічних тригерів, повторюваності повідомлень, спотворенні фактів і апеляції до емоцій.

Розглянула вплив дезінформації на суспільну думку, національну безпеку та інформаційну стійкість. Показано, що дезінформаційні кампанії здатні викликати паніку, недовіру до державних інституцій, посилення соціальної напруги та навіть впливати на політичні рішення.

Окремий акцент зробила на ролі соціальних мереж, ботів і фейкових акаунтів як ключових інструментів розповсюдження дезінформації в сучасному цифровому середовищі. Встановила, що швидкість, анонімність і широке охоплення аудиторії роблять ці платформи особливо вразливими до маніпуляцій та зовнішнього впливу.

Загалом, результати першого розділу створюють необхідне теоретичне підґрунтя для подальшого моделювання процесів поширення дезінформації та пошуку ефективних методів її нейтралізації.

РОЗДІЛ 2. АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ПРОТИДІЇ ДЕЗІНФОРМАЦІЇ

2.1. Методи автоматичного виявлення фейкових новин

Виявлення джерел дезінформації та фейкових новин є непростим завданням. Сьогодні запропоновано декілька підходів до розв'язання цієї проблеми. Одним із різновидів джерел дезінформації є фейкові новини, що являють собою свідоме подання неправдивої інформації. На основі аналізу можна виділити два основні підходи до виявлення джерел дезінформації: ручний та автоматичний. Для ручної перевірки фактів залучаються експерти. Автоматичні методи виявлення джерел дезінформації базуються на методах глибокого та машинного навчання. Більшість методологій поєднує декілька методів виявлення джерел дезінформації. На основі порівняння було виявлено, що методи, засновані на глибокому навчанні, є більш ефективними і використовуються багатьма дослідниками.

Класифікація відомих методологій для виявлення джерел дезінформації [22]:

Рис. 3. Класифікація відомих методологій для виявлення джерел дезінформації [22].

В автоматичних методах на основі глибокого навчання варто розшифрувати назви моделей, отже:

Модель RNN – рекурентна нейорнна мережа

Модель LSTM – довготривала короткочасна пам’ять

Модель Bi-LSTM – двонаправлена LSTM

Модель CNN – згорткова нейронна мережа

Тепер розглянемо детальніше ці автоматичні методи [23].

Методи на основі глибокого навчання.

Згорткова нейронна мережа (CNN) – це архітектура глибокого навчання, яка використовує згорткові шари для відображення ознак вхідних даних. Шари впорядковуються шляхом застосування фільтрів різних розмірів для створення різних відображень ознак. CNN може отримувати інформацію про вхідні дані на основі результатів відображення ознак. Шар об'єднання зазвичай супроводжує згортковий шар для отримання точних вихідних розмірів навіть з різними фільтрами. Шар об'єднання також полегшує обчислювальне навантаження, зменшуючи вихідні розмірності без втрати важливої інформації.

Архітектура CNN у цьому дослідженні була адаптована з дослідження Каліяра. Один вхідний шар з 1000 розмірностями з'єднаний з шаром вбудовування, розмірності якого визначаються розмірностями вбудовування попередньо навчених слів. Шар вбудовування з'єднаний з трьома одновимірними шарами згортки з різними розмірами ядра, де кожен шар згортки створюватиме різне відображення ознак. Кожен шар згортки з'єднаний з шаром максимального об'єднання для стиснення ознак та контролю перенавчання. Об'єднаний шар об'єднуватиме ознаки, отримані кожним шаром максимального об'єднання. Потім об'єднаний шар з'єднується з двома шарами згортки та шаром максимального об'єднання для подальшого вилучення ознак. Нарешті, глобальний шар максимального об'єднання з'єднаний з повністю зв'язаним шаром та вихідним шаром.

Двонаправлена LSTM (Bi-LSTM) - це архітектура типу рекурентної нейронної мережі (RNN), що складається з двох пристроїв з довгою короткочасною пам'яттю (LSTM), розташованих у різних напрямках. Архітектура спрямована на покращення можливостей пам'яті LSTM шляхом надання контекстної інформації з минулого та майбутнього.

Двонаправлена архітектуру LSTM використовується в цьому дослідженні, де є вхідний шар з 1000 вимірами, а потім шар вбудовування. Архітектура продовжується шаром двонаправленого LSTM, де LSTM ідентичний для обох напрямків. Крім того, глобальний шар максимального об'єднання, повністю зв'язаний шар та вихідний шар прогнозують клас [23].

Таблиця 1

Порівняння запропонованих методів з сучасним сучасним станом на основі набору даних про фейкові новини ISOT.

Автор	**Модель векторизації слів**	**Модель класифікації**	**Точність**	**Акуратність**	**Відгук**	**F1-оцінка**
Ahmed et al. [13]	TF-IDF	Linear SVM	92%	-	-	-
Ozbay & Alatas [14]	TF-IDF	Decision tree	96.8%	96.3%	97.3%	96.8%
Ahmad et al. [15]	LIWC	Random Forest	99%	99%	100%	-
Запропонована модель	fastText	CNN	99.88%	99.89%	99.88%	99.88%
Запропонована модель	GloVe	ResNet	99.90%	99.91%	99.90%	99.90%
Запропонована модель	GloVe	BidIrectional	99.95%	99.95%	99.95%	99.95%

Рекурентні нейронні мережі (RNN) – це тип штучної нейронної мережі, який часто використовується для аналізу послідовностей. Це означає, що вони спеціально розроблені для роботи з даними, що містять інформацію з часовим контекстом. RNN складаються з кількох шарів, і ці шари моделюються послідовно, щоб можна було моделювати послідовності слів та зв'язки між ними. RNN має

чудову здатність захоплювати контекстну інформацію з послідовностей слів, і ця інформація ефективно використовується для досягнення процесу класифікації даних. Цей тип мережі має основну функцію, яка полягає в частковому запам'ятовуванні того, що було на вхідних даних у контексті минулого часу. Це забезпечується завдяки існуванню прихованого стану та рекурентних зв'язків між певними нейронами або шарами. В результаті, наприклад, нейрони в прихованому шарі оснащені двома типами зв'язків. Окрім того, що вони з'єднані з іншим шаром, кожен нейрон також має рекурсивне з'єднання з самим собою.

Навчання рекурентної мережі може бути не таким простим, оскільки може виникнути проблема зникнення градієнта. Моделі навчання з кількома шарами можуть призвести до занадто малих або занадто великих градієнтів, оскільки помилки можуть передаватися через ці шари. Під час навчання на довших послідовностях слів ця проблема прискорюється, що означає, що навчання є складнішим. Рекурентні мережі добре працюють на коротких послідовностях, але для довших прогнозування наступної частини послідовності може вимагати інформації, отриманої майже на початку. Ця інформація часто втрачається в класичних рекурентних мережах. Довгострокова пам'ять (LSTM) та гейтовані рекурентні одиниці (GRU) можуть вирішити цю проблему пам'яті [22].

LSTM (довга короткострокова пам'ять) – це тип рекурентної нейронної мережі, яка специфічна своєю здатністю зберігати інформацію, що робить її основним кандидатом для зберігання інформації, яка в іншому випадку могла б загубитися в довшій послідовності слів. Базова структура LSTM складається з повторюваних модулів, відомих як блоки LSTM, які пов'язані один з одним та утворюють ланцюжок (див. Рисунок 3). Кожен блок містить 4 шари, і кожен шар включає окремі параметри мережі, які навчаються. Шари взаємодіють один з одним різними способами. В окремих комірках вентилі (вхідний вентиль, вентиль забуття та вихідний вентиль) використовуються для видалення або додавання інформації до стану блоку. Інформація, представлена у вигляді вектора слів, згодом проходить

через вентилі, що складаються з нейронів, оснащених сигмоїдною функцією активації. Блоки містять усі три типи вентилів, які регулюють стан, у якому знаходяться блоки на поточному кроці часу. Роль вентиля забуття полягає у вирішенні, чи слід забути передану інформацію, чи яку її частину слід позбутися [24].

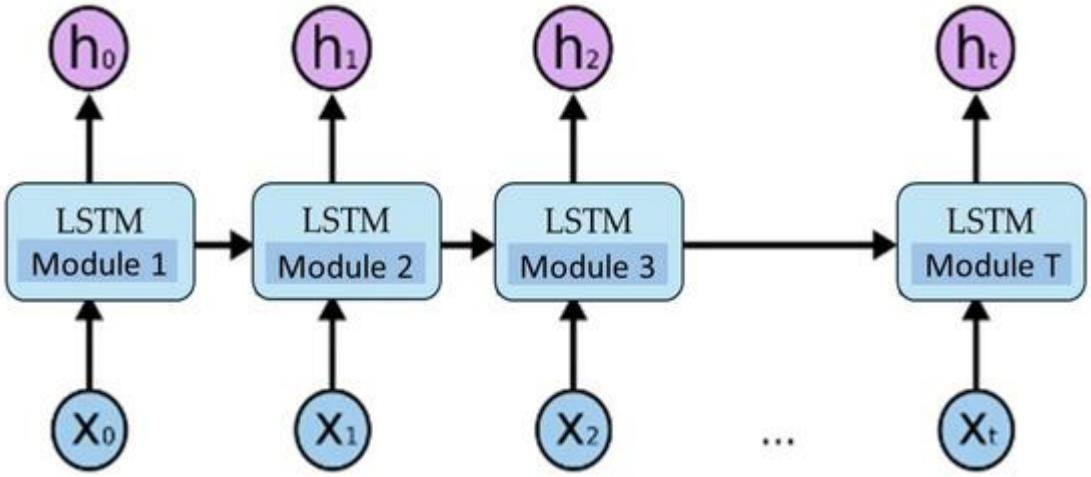

Рис. 3. Ілюстрація ланцюжка блоків LSTM [24]

Згідно Таблиці 1 метод Bi-LSTM, але розглянемо дослідження з приводу цього детальніше. У дослідженні [24] підготували навчальний набір даних та тестовий набір даних і розділили навчальні дані на три рівні частини для кожного клієнта-граничного елемента. Тільки тестовий набір даних використовувався для оцінки (точність, повнота, метрика F1) на глобальній моделі після кожного раунду навчання моделі.

Перша архітектура моделі базувалася на згортковій нейронній мережі (CNN). Функцією втрат була бінарна перехресна ентропія, а навчання було оптимізовано за допомогою оптимізатора Adam. Після кількох експериментів кількість прогонів було встановлено на 10. Кількість епох становила 1 для кожної моделі, а розмір пакета мав значення 25.

Перший шар цієї нейронної мережі – це Embedding, який представляє слова в тексті за допомогою вектора розміром 150. Ця архітектура містить кілька шарів згорткової мережі (Conv1D), які намагаються ідентифікувати різні шаблони в тексті за допомогою фільтрів розміром 5. Шари MaxPooling1D потім використовуються для зменшення розмірів після обробки попередніми згортковими шарами і таким чином зменшують кількість параметрів у мережі. Для покращення продуктивності та запобігання перенавчанню в моделі використовується шар Dropout зі значенням 0,3. Серед останніх шарів архітектури GlobalMaxPooling1D використовується для отримання одного значення для кожного фільтра. Зрештою, вихідний сигнал цього шару використовується для прогнозування єдиного двійкового виходу за допомогою щільних шарів з ReLU та сигмоїдними функціями активації. У таблиці 2 представлено результати тестування моделі CNN на основі топології без федеративного навчання, тобто модель була навчена на всіх нерозділених даних.

Таблиця 2

Ефективність моделі CNN з топологією, без застосування федеративного навчання

CNN	**Precision**	**Recall**	**F1 Rate**	**Support**
Real_news	0.97	0.97	0.97	814
Fake_news	0.89	0.88	0.89	208
Accuracy	-		0.95	1022
Macro average	0.93	0.93	0.93	1022
Weighted average	0.95	0.95	0.95	1022

Друга архітектура нейронної мережі базується на моделі з кількома LSTM. Ця топологія містить такі шари:

1. Шар вбудовування, який перетворює вхідний текст на навчені вбудовування слів за допомогою алгоритму word2vec;
2. Шар LSTM, що містить 64 нейрони з функцією активації Tanh, за допомогою якої модель може зберігати довгострокові залежності між словами вхідного тексту;

3. Шар відсіву, що використовується для запобігання перенавчанню моделі та видалення деяких випадкових виходів з попереднього шару;
4. Другий шар LSTM, що містить 32 нейрони з функцією активації Tanh, що покращує збереження довгострокових залежностей слів тексту;
5. Щільний шар відсіву, що використовується для видалення випадкових виходів з попереднього шару;
6. Щільний шар, подібний до попереднього, що містить 16 нейронів з функцією активації ReLU;
7. Щільний шар як останній вихідний шар, що містить один нейрон із сигмоподібною функцією активації, яка використовується для двійкової класифікації виходу моделі.

Під час розробки цієї моделі вони вирішили використовувати бінарну функцію втрат крос-ентропії. Для оптимізації вони обрали алгоритм Адама, і після ітеративного тестування вирішили навчати модель протягом 10 раундів.

У Таблиці 3 наведено результати тестування моделі з кількома LSTM на основі топології з таблиці 5 без федеративного навчання, тобто модель була навчена на всіх нерозділених даних.

Таблиця 3

Ефективність множинної моделі LSTM, без застосування федеративного навчання

LSTM + LSTM	**Precision**	**Recall**	**F1 Rate**	**Support**
Real_news	0.97	0.99	0.98	814
Fake_news	0.92	0.85	0.88	208
Accuracy	-		0.97	1022
Macro average	0.95	0.93	0.94	1022
Weighted average	0.96	0.96	0.96	1022

Третя архітектура нейронної мережі базується на моделі CNN та моделі LSTM. Ця архітектура складається з наступних шарів:

1) Шар вбудовування для перетворення вхідного тексту у форму навчених вбудовувань слів за допомогою word2vec;

2) Шар Conv1D, що містить 32 фільтри з розміром вікна 3 та функцією активації ReLU, яка забезпечує вилучення ознак з вхідного тексту;
3) Шар Dropout, що запобігає попередньому навчанню моделі та видаляє деякі випадкові виходи з попереднього шару;
4) Шар MaxPooling1D, що зменшує розмірність виходів з згорткового шару;
5) Шар LSTM, що містить 32 нейрони та функцію активації Tanh і дозволяє навчальній моделі зберігати довгострокові залежності між словами вхідного тексту;
6) Щільний шар Dropout, що видаляє деякі випадкові виходи з попереднього шару;
7) Підлий шар як передостанній шар, що містить 8 нейронів з функцією активації ReLU;
8) Щільний шар як вихідний шар, що містить 1 нейрон із сигмоподібною функцією активації, яка використовується для бінарної класифікації виходу моделі.

Як функцію втрат використовувалася бінарна перехресна ентропія. Вони використовували алгоритм оптимізації Адама, і після кількох експериментів вони вирішили провести 10 раундів навчання моделі.

У Таблиці 4 наведено результати тестування моделі CNN + LSTM на основі топології з таблиці 3 без федеративного навчання, тобто модель була навчена на всіх нерозділених даних.

Таблиця 4

Ефективність моделі CNN + LSTM без застосування федеративного навчання

CNN + LSTM	**Precision**	**Recall**	**F1 Rate**	**Support**
Real news	0.96	0.98	0.97	814
Fake news	0.91	0.86	0.88	208
Accuracy	-		0.96	1022
Macro average	0.94	0.92	0.93	1022
Weighted average	0.96	0.96	0.96	1022

Результати експериментів, представлені в Таблиці 2, Таблиці 3 та Таблиці 4, показують, що найефективнішою моделлю була друга архітектура з кількома шарами LSTM, а найменш ефективною – перша архітектура CNN, коли навчання проводилося на всьому нерозділеному наборі даних без застосування федеративного навчання [24].

Федеративне навчання

Сучасні розподілені мережі та пристрої щодня генерують великі обсяги різних даних. З їх постійно зростаючою потужністю та можливостями, користувачі можуть мати занепокоєння щодо передачі приватних даних та інформації. Наприклад, оскільки можливості зберігання даних та обчислювальна потужність пристроїв у розподілених мережах множаться, використання локальної ємності сховища на кожному локальному вузлі не є проблемою; таким чином, захист конфіденційності під час передачі конфіденційних даних не є необхідним, оскільки дані, створені користувачем, залишаються на локальних пристроях. Тому нещодавня перевага локальному зберіганню даних та моделям навчання локально віддає перевагу федеративному навчанню, яке дозволяє навчати моделі безпосередньо на віддалених пристроях. Федеративне навчання є можливим рішенням, оскільки воно може досягти зменшення витоку даних. Розробка цієї технології має потенціал перетворити сучасні інтелектуальні системи Інтернету речей з передовими архітектурами федеративного навчання на системи, що підвищують конфіденційність. Федеративне навчання – це форма розподіленого навчання моделей. Замість навчання на центральній кінцевій точці, навчання виконується на кількох клієнтах перед агрегацією. Федеративне навчання використовувалося в багатьох різних пристроях, наприклад, для класифікації медичних зображень та інших структурованих даних у налаштуваннях пристроїв Інтернету речей.

Вони класифікують федеративне навчання на два типи на основі мережевої структури:

- Централізоване федеративне навчання — найпоширеніший підхід до федеративного навчання. Цей підхід базується на центральному сервері та наборі клієнтів. У кожному циклі навчання всі клієнти забезпечують паралельне навчання моделей нейронної мережі, використовуючи свої локально отримані набори даних. Відповідно, всі клієнти передають набори параметрів моделей на центральний сервер для агрегації, використовуючи, наприклад, алгоритм зваженого усереднення. На центральному сервері створюється глобальна модель та надсилається клієнтам для повторного навчання агрегованої моделі на їхніх локальних даних. Нарешті, коли процес навчання завершено, кожен клієнт має у своєму розпорядженні ту саму глобальну модель та персоналізовану версію своєї локальної моделі. Тут центральний сервер вважається ключовим компонентом цього виду федеративного навчання для централізованого агрегування локальних моделей та розповсюдження глобальної моделі клієнтам.
- Децентралізоване федеративне навчання — мережева топологія, на відміну від централізованого федеративного навчання, де всі клієнти спілкуються між собою рівноправним способом. Таким чином, клієнти навчають локальні моделі на своїх наборах даних у кожному раунді зв'язку. Після цього кожен клієнт оновлює моделі, отримані від сусідніх клієнтів, через одноранговий зв'язок, щоб досягти консенсусу щодо глобального оновлення. Децентралізоване федеративне навчання призначене для повної або часткової заміни центрального сервера, якщо, наприклад, зв'язок із сервером недоступний.

Федеративне навчання підвищує конфіденційність даних шляхом навчання моделей машинного навчання на локальних приміщень, де знаходяться дані, тому передача даних не потрібна. Тоді воно дозволяє ділитися лише результуючою

моделлю. Таким чином, таке навчання дозволяє уникнути необхідності обміну даними між локальними станціями – клієнтами.

Паралельне федеративне навчання – це тип централізованого федеративного навчання моделей у розподіленому середовищі. Для цієї методики моделі локальних нейронних мереж навчаються на кожній моделі на межі клієнта на першому кроці. Потім ваги, отримані з локальних моделей, усереднюються та тестуються на глобальній моделі в агрегаторі. Потім ваги, отримані з глобальної моделі, повертаються до локальних моделей, і нейронна мережа знову навчається там. Цей цикл повторюється, доки не буде досягнуто заданої кількості раундів навчання.

В експерименті вони використовували паралельне федеративне навчання та застосували його через бібліотеку Ray. У їхньому рішенні вони моделювали федеративне навчання, використовуючи Ray як фреймворк для розподіленої обробки завдань, та створили децентралізовану мережу клієнтів на одній машині, де кожен клієнт мав власні дані. На початку методу необхідно було визначити кількість раундів для навчання, тобто скільки разів застосовувати повний цикл федеративного навчання. На рисунку 4 показано паралельний цикл федеративного навчання з трьома ребрами – клієнтами.

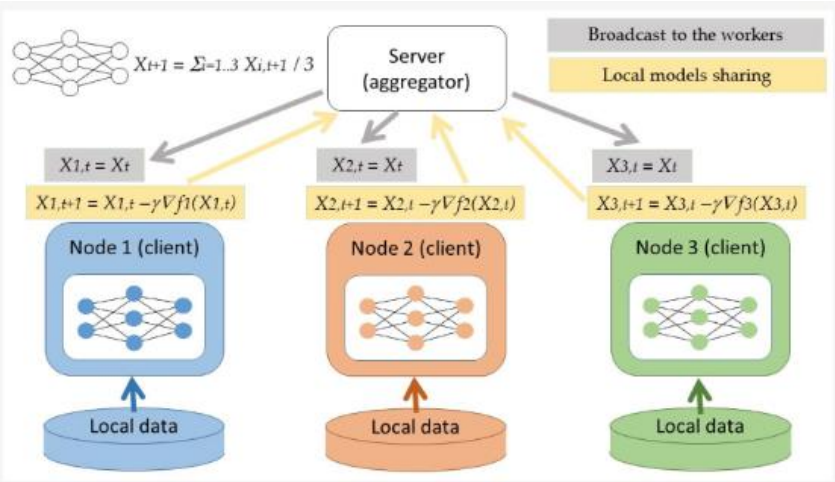

Рис. 4**.** Архітектурна схема паралельного федеративного навчання з трьома ребрами-клієнтами

Як ми бачимо на рисунку 4, сервер та всі клієнти мають однакову структуру моделі – одну з вищезгаданих архітектур моделей (CNN, CNN + LSTM або кілька LSTM). Сервер ініціалізує свою копію моделі (глобальну модель) та запускає певну кількість раундів навчання. Усі раунди проводяться з однаковою структурою. У раунді, що починається з моменту часу t, сервер надає клієнтам спільну інформацію про свою глобальну модель (тобто вектор ваг X_t). Кожен клієнт використовує ваги для ініціалізації своєї локальної копії моделі ($X_{i,t} = X_t$) та виконує одну епоху навчання, використовуючи свої локально доступні дані (за допомогою алгоритму, заснованого на градієнтному спуску), для створення локальної моделі. Згодом клієнти діляться своїми локальними моделями (тобто ваговими векторами $X_{i,t+1}$) з сервером. Нарешті, сервер агрегує локальні моделі у вигляді нової глобальної моделі X_{t+1}.

Застосування федеративного навчання до трьох локальних наборів даних з використанням розроблених топологій

Результати експериментів з федеративним навчанням на першій моделі глибокого навчання з архітектурою CNN представлені в Таблиці 5. Три моделі CNN були навчені для трьох клієнтів, а потім агреговані на центральному сервері.

Таблиця 5

Ефективність федеративного навчання моделей CNN, навчених на трьох локальних наборах даних на локальних сторонах трьох клієнтів.

CNN	Percision	Recall	F1 Rate	Support
Client 1				
Real_news	0.99	0.93	0.96	814
Fake_news	0.77	0.96	0.86	208
Accuracy			0.93	1022
Macro average	0.88	0.94	0.91	1022

Продовження таблиці 5

Weighted average	0.95	0.93	0.94	1022
Client 2				
Real_news	0.98	0.97	0.97	814
Fake_news	0.89	0.91	0.90	208
Accuracy			0.96	1022
Macro average	0.93	0.94	0.93	1022
Weighted average	0.96	0.96	0.96	1022
Client 3				
Real_news	0.95	0.99	0.97	814
Fake_news	0.95	0.78	0.86	208
Accuracy			0.95	1022
Macro average	0.95	0.88	0.91	1022
Weighted average	0.95	0.95	0.95	1022

У таблиці 6 представлено результати кінцевої моделі CNN після всіх циклів паралельного федеративного навчання. Кінцева модель є результатом агрегації на центральному сервері.

Таблиця 6

Ефективність кінцевої агрегованої моделі CNN, навченої з використанням паралельного федеративного навчання

CNN	**Precision**	**Recall**	**F1 Rate**	**Support**
Real_news	0.98	0.97	0.97	814
Fake_news	0.88	0.93	0.90	208
Accuracy			0.96	1022
Macro average	0.93	0.95	0.94	1022
Weighted average	0.96	0.96	0.96	1022

У наступних двох таблицях представлено результати агрегованих моделей, навчених за допомогою паралельного федеративного навчання. У таблиці 7 показано ефективність агрегованої моделі з кількома LSTM, а в таблиці 8 показано ефективність агрегованої моделі CNN + LSTM.

Таблиця 7

Ефективність агрегованої множинної LSTM-моделі, навченої за допомогою паралельного федеративного навчання

CNN	Precision	Recall	F1 Rate	Support
Real_news	0.96	0.97	0.97	814
Fake_news	0.89	0.85	0.87	208
Accuracy			0.95	1022
Macro average	0.93	0.91	0.92	1022
Weighted average	0.95	0.95	0.95	1022

Таблиця 8

Ефективність агрегованої моделі CNN + LSTM, навченої з використанням паралельного федеративного навчання

CNN	Precision	Recall	F1 Rate	Support
Real_news	0.97	0.97	0.97	814
Fake_news	0.88	0.89	0.89	208
Accuracy			0.95	1022
Macro average	0.93	0.93	0.93	1022
Weighted average	0.95	0.95	0.95	1022

Результати експериментів, представлені в Таблиці 6, Таблиці 7 та Таблиці 8, демонструють протилежну тенденцію до результатів, наведених для всього нерозділеного набору даних без застосування федеративного навчання. При використанні федеративного навчання найефективнішою моделлю була перша

модель з архітектурою CNN, а найменш ефективною – друга архітектура, що базується на кількох шарах LSTM, хоча відмінності в значеннях швидкості F1, точності та повноти незначні [24].

Метод опорних векторів з радіальною базисною функцією (SVM-RBF), відомий алгоритм класифікації, використовує трюк ядра для відображення даних у багатовимірний простір, що полегшує ідентифікацію гіперплощини, яка розділяє класи. Для цієї мети використовується ядро RBF.

$$\text{Мінімізувати } \sum i = 1\ t_0\, N\ \alpha i - \frac{1}{2} \sum i = 1 t_0 N \sum j = 1\ t_0\, N\ \alpha i\ \alpha j\ yi * yj\ K(xi, xj) \quad (1)$$

де:

N = кількість точок даних

αi, αj = множники Лагранжа

yi, yj = цільові значення

xi, xj = вхідні вектори

K(xi, xj) = функція ядра RBF

Алгоритм SVM-RBF визначає оптимальну гіперплощину, максимізуючи запас між гіперплощиною та найближчими точками даних. Він використовує ядро RBF для відображення даних у багатовимірний простір для покращеного розділення.

Під час сортування даних за різними категоріями мало які методи є такими ж ефективними, як алгоритм опорних векторів (SVM). Програма починається зі зчитування позначених випадків з набору даних, як показано в алгоритмі 1. Потім дані розділяються на навчальний набір та тестовий набір. Функцію ядра RBF необхідно визначити за допомогою навчальних даних для навчання класифікатора SVM. Визначення граничних значень між категоріями неможливе без обчислення подібності між точками даних функцією ядра RBF. Під час фази навчання

класифікатор SVM використовує навчальний набір та функцію ядра RBF для визначення найкращої гіперплощини, яка оптимізує межу між класами. Після навчання класифікатора його ефективність вимірюється на тестовому наборі. Точність класифікатора SVM можна визначити шляхом порівняння передбачуваних міток класів з фактичними мітками класів тестового набору. Під час роботи з нелінійними межами прийняття рішень та високовимірними даними метод SVM часто використовується для задач класифікації. Його універсальність та точність зробили його найкращим засобом для застосувань, починаючи від розпізнавання зображень та категоризації тексту до біоінформатики.

На рисунку 5 показано запропоновану блок-схему розробленої системи. Видно, що система має чотири основні компоненти: набір даних, компонент SVM-RBF, миттєву KNN та результати. Набір даних спочатку завантажується в систему, після чого над даними виконуються такі методи попередньої обробки, як нормалізація, соціальна інженерія, масштабування тощо. Потім набір даних розділяється на навчальний та тестовий. Навчальний набір даних використовується для побудови SVM-RBF та миттєвої KNN, а тестовий набір даних – для оцінки побудованої моделі за допомогою матриці плутанини, та визначення таких показників продуктивності, як точність, прецизійність, повнота тощо.

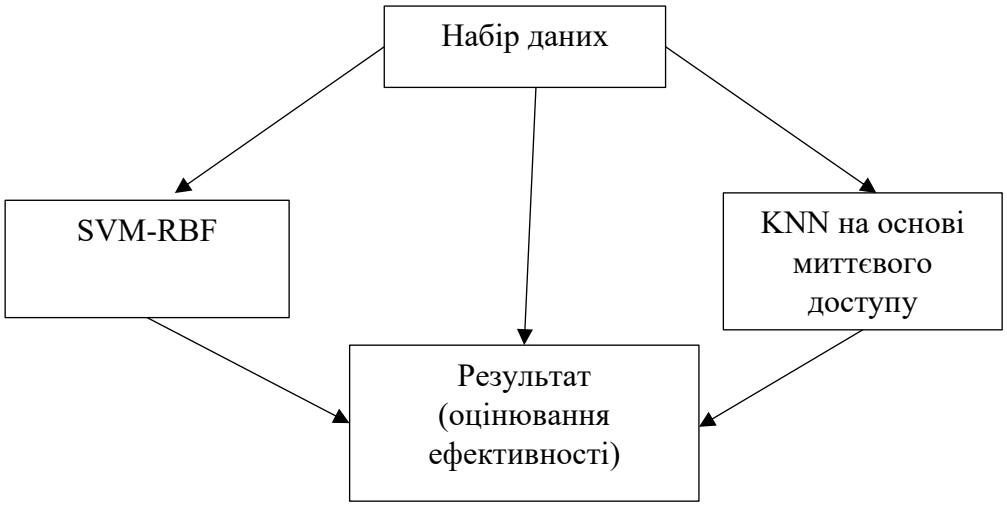

Рис. 5. Запропонований дизайн системи

Метод SVM-RBF використовується для класифікації, оскільки він добре працює в різних програмах машинного навчання. Його здатність працювати з високовимірними даними та нелінійними кореляціями робить його ідеальним інструментом для виявлення фальшивих новин. Метод KNN використовується для вибору ознак, оскільки він ефективно визначає, які характеристики найбільш релевантні для проблеми класифікації. Ефективність можна підвищити, ретельно вибравши відповідні елементи для проблеми.

Продуктивність запропонованої системи можна ретельно оцінити за допомогою ретельного експериментування та порівняльного аналізу. Це доводить перевагу та ефективність запропонованої стратегії у виявленні та протидії фейковим новинам шляхом порівняння результатів із сучасними підходами. Запропоноване рішення використовує передові методи машинного навчання, спрощує вибір ознак та гарантує точні рейтинги, поєднуючи ці елементи. Точність системи у виявленні фейкових новин покращується завдяки цьому всеохоплюючому методу, що робить її потужним та надійним інструментом для виявлення та протидії дезінформації в електронній пошті.

Таблиця 9.

Порівняльний аналіз запропонованих моделей

Оцінювання ефективності	**SVM-RBF**	**IB-KNN**
Точність	97.42	91.21
Чутливість	99.39	94.54
Специфіка	91.19	80.77
Точність позитивного предбачення	97.28	93.92
Оцінка F1	98.32	94.23
Збалансований коефіцієнт класифікації	95.29	87.66

Це дослідження довело, що методи машинного навчання, особливо SVM-RBF та IB-KNN, можуть ефективно виявляти фальшиві пересилання електронної пошти. Результати показали, що SVM-RBF має вищу точність (97%), ніж IB-KNN (85%). Тим не менш, як використані набори даних, так і алгоритми класифікації можуть бути вдосконалені. У майбутніх дослідженнях слід дослідити більш повні та різноманітні набори даних, такі як дані з різних сайтів соціальних мереж. Це дозволить провести більш ретельний аналіз ефективності класифікаторів на ширшому діапазоні даних. Подальше покращення продуктивності може бути можливим шляхом вивчення та розробки гібридних алгоритмів, які поєднують переваги різних класифікаторів, таких як SVM-RBF та IB-KNN. Точність класифікації електронної пошти та запобігання передачі фейкових новин та іншого небезпечного контенту можна підвищити шляхом інтеграції заздалегідь визначених каталогів у поштові сервери та програми, таких як користувацькі та контекстно-залежні довідники. На завершення, це дослідження показало, що методи машинного навчання, особливо SVM-RBF та IB-KNN, допомагають виявляти містифікації в електронній пошті. Тим не менш, необхідні вдосконалення в розробці алгоритмів та доступності даних. Майбутні дослідження в цих сферах допоможуть покращити точність та ефективність класифікації, що сприятиме боротьбі з фейковими новинами та захищатиме людей від дезінформації [25].

Всі інші методи основані на машинному навчанні коротко описані нижче.

Мультиноміальний наївний баєсівський класифікатор, названий так тому, що це баєсівський класифікатор, який робить спрощене (наївне) припущення про те, як взаємодіють ознаки.

Інтуїтивне розуміння класифікатора показано на рис. 6.

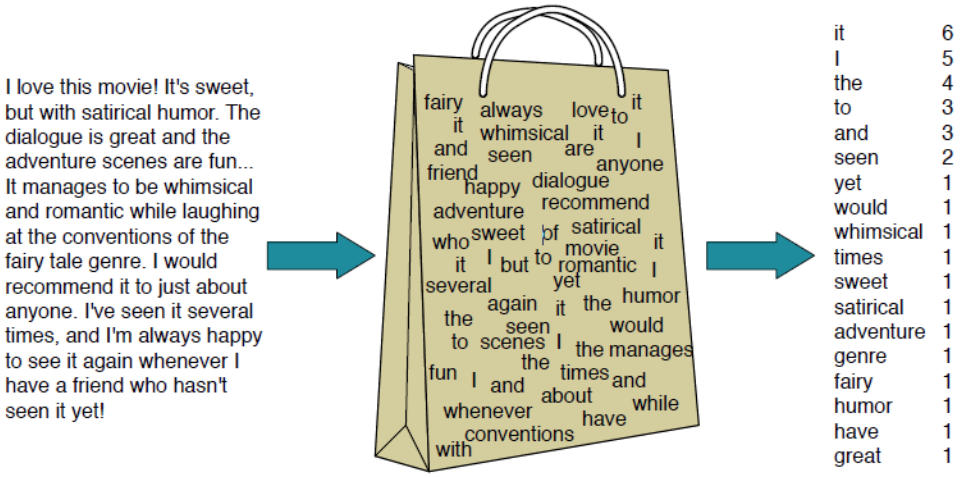

Рис. 6**.** Логіка наївного баєсівського класифікатора

Ми представляємо текстовий документ як мішок слів, тобто невпорядкований набір слів, позиція яких ігнорується, зберігаючи лише їх частоту в документі. У прикладі на рисунку, замість того, щоб представляти порядок слів у всіх фразах, таких як «Я люблю цей фільм» та «Я б рекомендував його», ми просто зазначаємо, що слово «я» зустрічалося 5 разів у всьому уривку, слово «оно» – 6 разів, слова «люблю», «рекомендую» та «фільм» – один раз тощо.

Спираючись на приклад використання цього методу можна зробити висновки, що така модель потребує великий об’єм даних та використання двійкової класифікації (наприклад «фейк» чи «не фейк»), тому що якщо класів більше це призводить до неоднозначних рішень [26].

Випадковий ліс – метод в якому набір даних випадково розділяється на зразки і на кожному кроці дерева обираються не весь набір ознак, а лише випадкову його частину. У фінальному рішенні для класифікації використовується голосування більшості, а для регресії – середнє значення [27].

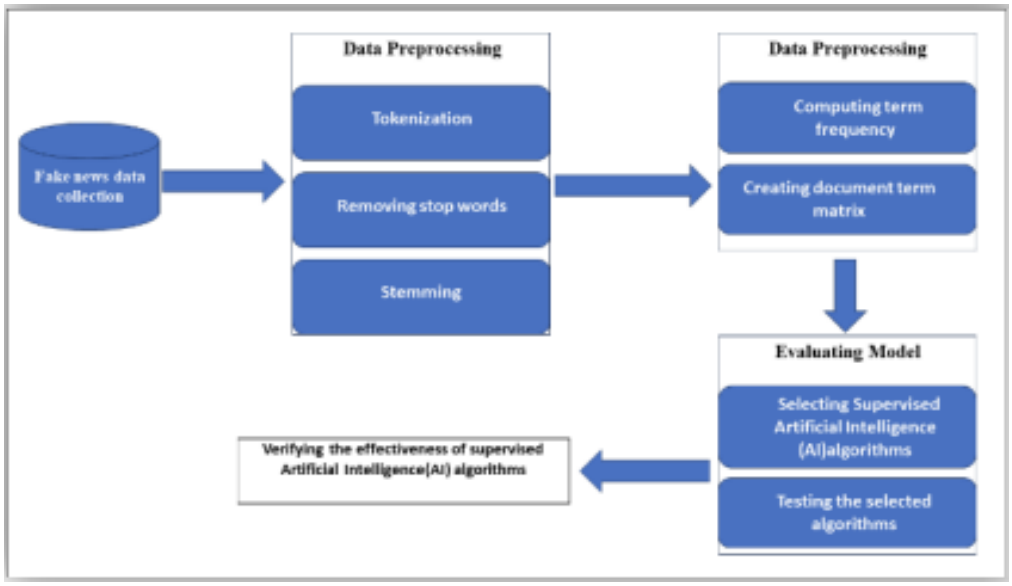

Рис. 7. Модель випадкового лісу [27]

Дерева рішень – метод, в якому усі дані дослідження повинні бути або числовими, або категоріальними за своїм видом. Він використовує два різних методи обрізання.

- Заміна піддерева – це перший метод, який вказує на потенціал заміни окремих листків дерева рішень, щоб зменшити загальну кількість тестів вздовж переконливого маршруту.
- Підвищення піддерева часто має незначний вплив на моделі дерев рішень.

Алгоритм починається з кореневого вузла. Він перевіряє умову, наприклад: “чи значення ознаки X > 5?” Якщо «так» — йдемо однією гілкою, якщо «ні» — іншою. Цей процес повторюється, доки ми не дійдемо до листка дерева, де приймається рішення [27].

Логістична регресія – це метод статистичного аналізу для прогнозування бінарного результату, такого як «так» чи «ні» (бінарна класифікація), на основі попередніх спостережень набору даних. Це метод контрольованої статистики для знаходження ймовірності залежної змінної [28]. Її назвали на честь функції, яка

використовується в основі методу, – логістичної функції. Логістична функція, яку також називають сигмоподібною функцією, була розроблена статистиками для опису властивостей зростання популяції в екології, швидкого зростання та досягнення максимуму на межі несучої здатності середовища. Це S-подібна крива, яка може прийняти будь-яке дійсне число та відобразити його у значення від 0 до 1, але ніколи не точно в цих межах [28].

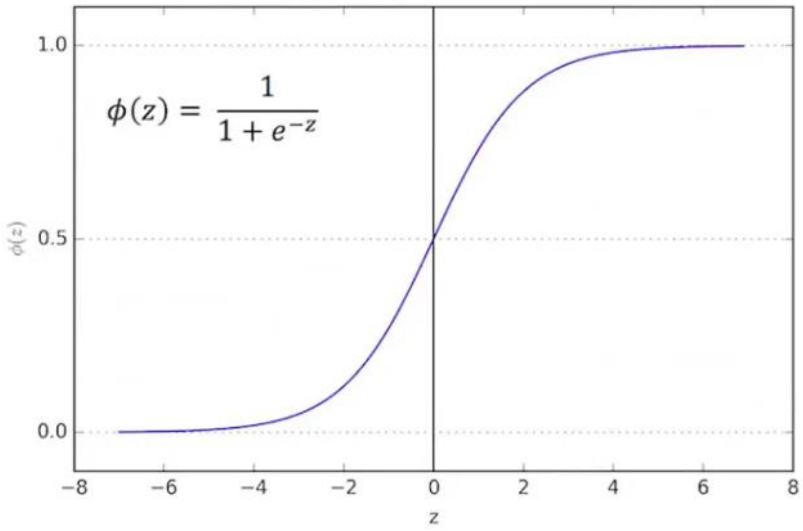

Рис. 8. Сигмоїдна функція [29]

Баєсівське моделювання використовує теорію ймовірностей для виявлення та аналізу закономірностей у даних і формулювання висновків з них. Воно добре підходить для завдання виявлення фейкових новин, оскільки його можна використовувати для визначення ймовірності того, що новини є неправдивими, на основі їхнього змісту. Наприклад, якщо новина, здається, містить неправдиву інформацію, вона може бути пов'язана з певними фразами або словами, які зазвичай використовуються у фейкових новинах. За допомогою баєсівського моделювання ці закономірності можна ідентифікувати та використовувати для розробки моделі, яка ідентифікує та класифікує історії. Крім того, баєсівські апріорні версії базуються на ймовірності настання події, що в контексті фейкових новин дозволяє дослідникам вимірювати, наскільки ймовірно, що новина є правдивою. Наприклад,

якщо новину повідомляє відоме ЗМІ, дослідники можуть призначити вищу ймовірність того, що ця новина є правдивою, ніж якби її повідомило невідоме джерело. Аналогічно, коли кілька надійних джерел повідомляють про новину, дослідники можуть призначити вищу ймовірність того, що новина є правдивою, ніж якби її повідомило лише одне джерело. Багато методів моделювання, включаючи алгоритми глибокого навчання, що використовують частоти або стандартні концепції умовної ймовірності, не можуть скористатися такою априорною інформацією чи знаннями. Більше того, баєсівські методи можна використовувати для оцінки оптимального розподілу нейронної мережі для фейкових новин. Використовуючи баєсівську стратегію, ми можемо бути впевнені в нашій гіпотезі або проти неї, а не приходити до єдиного висновку [30].

2.2. Використання штучного інтелекту для аналізу інформації

Переваги та вигоди від використання інструментів на основі штучного інтелекту для аналізу даних [31]:

- Доступність: Ці інструменти розроблені з урахуванням потреб користувачів без технічних знань, мають інтуїтивно зрозумілі інтерфейси та можливості обробки природної мови.
- Універсальність: Вони можуть обробляти різні формати даних, від електронних таблиць Excel до звітів PDF, що робить їх адаптованими до ваших існуючих робочих процесів.
- Швидкість: Алгоритми штучного інтелекту можуть обробляти та аналізувати великі набори даних набагато швидше, ніж традиційні методи, забезпечуючи швидке отримання аналітичних даних.
- Економічно ефективність: Багато з цих інструментів пропонують доступні тарифні плани, що робить розширений аналіз даних доступним для підприємств будь-якого розміру

Ось кроки для аналізу даних за допомогою ШІ [31]:

Оберіть правильний інструмент: Кілька платформ на базі штучного інтелекту (ШІ) задовольняють різні потреби та рівні навичок. Такі інструменти, як Ajelix, Promptloop та Numberry AI, спеціалізуються на аналізі та автоматизації Excel, дозволяючи маніпулювати даними за допомогою простих команд природною мовою. MonkeyLearn ідеально підходить для аналізу тексту для будь-якої таблиці Google. Це безкоштовне доповнення може отримувати аналітичні дані з опитувань, відгуків клієнтів та PDF-файлів з великим вмістом тексту. Klipfolio – це доступний та універсальний інструмент для аналізу та візуалізації даних, який легко інтегрується з Excel та іншими поширеними форматами та ідеально підходить для створення інтерактивних інформаційних панелей. Sheet AI безпосередньо переносить можливості ШІ в Google Таблиці, дозволяючи користувачам запитувати свої дані, використовуючи звичайну англійську мову.

Підготовка даних: Перш ніж розпочати, переконайтеся, що ваші дані організовані та чисті. Для таблиць Excel використовуйте чіткі заголовки та послідовне форматування. Під час роботи з PDF-файлами зосередьтеся на документах зі структурованими даними, такими як звіти чи рахунки-фактури.

Завантаження та аналіз: Після вибору інструменту завантажте свої дані та почніть досліджувати. Багато інструментів ШІ дозволяють ставити запитання про ваші дані природною мовою. Наприклад, ви можете запитати: «Які наші продукти були найпопулярнішими за останній квартал?» або «Підсумуйте ключові моменти з цього звіту про відгуки клієнтів». Використовуйте агентів штучного інтелекту для проведення аналізу за вас.

Візуалізація та інтерпретація: Дозвольте інструменту штучного інтелекту допомогти вам у створенні візуальних представлень ваших даних. Це можуть бути діаграми, графіки або інтерактивні панелі інструментів. Приділіть час, щоб дослідити різноманітні візуалізації, які найкраще підходять для вас, шукаючи

закономірності, тенденції або аномалії, які можуть вплинути на ваші бізнес-рішення.

Діліться та дійте: Багато інструментів штучного інтелекту пропонують прості варіанти обміну, що дозволяє вам співпрацювати з вашою командою або представляти результати зацікавленим сторонам. Використовуйте ці дані для прийняття обґрунтованих рішень у вашій організації.

Також існує кілька порад, щоб покращити результати [31]:

1) Робота з агентами штучного інтелекту вимагатиме кількох ітерацій, перш ніж бажаний результат відповідатиме тому, що надає інструмент штучного інтелекту. Пам’ятайте, що агенту потрібен контекст та конкретні інструкції. Чим краще ви обробляєте інструкції, тим кращий результат, і це вимагатиме певної практики.

2) Почніть з невеликого набору даних, з яким ви знайомі, пропустіть його через агента штучного інтелекту та переконайтеся, що результати відповідають тому, що ви знаєте про дані. Як тільки ви звикнете, переходьте до більшого набору даних.

3) Людина в циклі є ключовою для ШІ. Це означає, що незалежно від результату аналізу, завжди перевіряйте роботу агента штучного інтелекту. Ніколи не пропускайте перевірку фактів і перевіряйте результати під тиском, як ви робили б це з молодшим співробітником [31].

2.3. Фактчекінг та алгоритмічні підходи до перевірки достовірності новин

Для того, щоб розпізнати дезінформацію та запобігти її поширенню варто притримуватись деяких правил:

1. Створюйте різноманітну екосистему новинних медіа

Отримання новин з різних джерел та видання дозволяє широкому спектру інформації потрапляти до вашого звичайного новинного контенту. Якщо та сама

новина підтверджується кількома каналами, є більша ймовірність того, що це не дезінформація.

Швидкий пошук у Google може допомогти вам визначити, чи обговорюють цю тему інші надійні джерела. Якщо ні, ймовірність того, що це фейкові новини, значно зростає.

2. Критично оцінюйте достовірність джерел

Якщо ви сумніваєтеся, виберіть, яка інформація потребує подальшого розслідування або додаткових питань. З інформацією, яка надходить із соціальних мереж, відстеження інформації до оригінального джерела – чудовий спосіб боротьби з дезінформацією. Якщо інформація надходить з менш авторитетного джерела або суперечить порадам експертів, скептично ставтеся до представленої інформації та зауважте, що канал, де ви знайшли інформацію, може бути не найкращим.

3. Посилюйте голоси експертів

Підтримуйте справжніх експертів, які є видатними у своїй галузі. З будь-яким контентом, після його отримання, важливо дізнатися про кваліфікацію автора. Чи має цей експерт науковий ступінь, пов'язаний з цією темою? Чи має цей експерт багаторічний досвід роботи у своїй галузі? Чи вважають його авторитетом інші люди в його галузі? Чи є ця інформація рецензованою з авторитетного академічного журналу?

4. Спростуйте чутки та поясніть, чому вони невірні

Як тільки ви отримаєте факти, будьте прямолінійними та лаконічними. Уникайте наголосу на дезінформації, коли спростовуєте неправдиві твердження. Пояснення, чому дезінформація є невірною, ефективніше, ніж просто називати її хибною [32].

2.4. Аналіз державних та міжнародних ініціатив з боротьби з дезінформацією

Під час боротьби з дезінформацією глобальна спільнота використовує підхід, в якому дезінформація сприймається як загроза правам людини (наприклад, для прав на свободу голосу під час виборів, свободу переконань, совісті та релігії, повагу до особистого життя тощо). У той же час протидія дезінформації регулюється більш у рамках м'якого права, тобто націлена на пошук її першопричин, модерування контенту, а не його масове блокування та насамперед притягнення до відповідальності осіб. Зважаючи на це протягом останніх років у галузі боротьби з дезінформацією були прийняті певні стандарти. У Резолюції [33] Парламентської асамблеї Ради Європи № 2326 "Демократія зламана? Як відповісти?" написано про необхідність вдосконалення процесів у мережі Інтернет, які б дозволили боротися з дезінформацією, а також про посилення співпраці і інвестування у якісну журналістику.

Кодекс [34] практики ЄС з протидії дезінформації зобов'язує підписантів розробляти політики для протидії недобросовісній рекламі, зокрема, політичній, неправомірному використанню ботів. Також йдеться про необхідність оприлюднення інформації щодо замовників політичної і тематичної реклами, інформації про те, чому ця реклама спрямована на певну особу, покращення пошукових можливостей для користувачів тощо.

CM/Rec(2018)2 [35] рекомендує посилювати відповідальність провайдерів (надавачів послуг) за порушення прав їхніх клієнтів. Зокрема, покладає обов'язок модерувати контент, контролювати захист персональних даних, та загалом гарантувати дотримання прав людини у мережі. Також ухвалено Рекомендацію CM / Rec (2020) 1 Комітету Міністрів державам-членам щодо впливу алгоритмічних систем на права людини. Тому що алгоритмічні системи застосовуються для

розповсюдження цільової реклами, в тому числі і політичної, що може нести небезпеку поширення дезінформації.

Також країни впроваджують політики і законодавство щодо протидії дезінформації на національному рівні. Згідно з певними досліженнями [36] Фінляндія, вважається найбільш стійкою до неправдивої інформації, впровадила [37] навчальні курси щоб допомогти школярам виявляти дезінформацію, розвивати критичне мислення та використовувати фактчекінг.

У 2018 році Франція ухвалила закон “Проти маніпуляції інформацією” [37]. Він, дозволяє суддям ідентифікувати фейкові новини та припиняти їхнє поширення. Французькому організації мовлення можуть блокувати телевізійні медіа, які контролюються з-за кордону і діють проти державних інтересів.

Впровадження вищезазначених практик і ухвалення нормативно-правових актів демонструють свою ефективність. Так, Facebook кілька років тому почав розкривати замовників політичної реклами, ретельніше видаляти фейкові акаунти, виявляти скоординовану поведінку у соцмережі, спрямовану на поширення дезінформації. Про це компанія прозвітувалася [38] у повідомленні щодо імплементації Кодексу практик ЄС із протидії дезінформації.

В Україні немає юридичного визначення дезінформації. Не зважаючи на те, що у Законі України “Про інформацію” [39] одним із принципів інформаційних відносин визначається достовірність та цілісність інформації, а в законах “Про друковані засоби масової інформації (пресу) в Україні” [40] і “Про телебачення та радіомовлення” [41] йдеться про те, що журналісти та творчі працівники зобов’язані подавати об’єктивну і достовірну інформацію, попередньо перевіряючи її, жодного регулювання дезінформація як явище не має.

Закон про ЗМІ в Україні передбачає відповідальність за зловживання правом на інформацію: заклики до повалення конституційного ладу, розпалювання ворожнечі, пропаганди винятковості, зверхності тощо. Ці процеси звісно можуть

вчинятися і за допомогою дезінформації, наприклад, проросійськими ЗМІ, але вони не мають ознак перелічених вище правопорушень.

Насправді єдиним механізмом спростування інформації є вимога конкретної особи, про яку було здійснено наклеп або такі, які порушують право на честь, гідність і ділову репутацію.

Міністерство культури та інформаційної політики розробило проєкт закону "Про протидію дезінформації" [42], який регулює дезінформацію та закликає до відповідальності. У ньому юридично визначалося поняття дезінформації, впроваджувалася відповідальність за її поширення, утворювалися нові інституції. Однак законопроєкт так і не було зареєстровано у Верховній Раді через те, що деякі його положення були піддані критиці за надміру жорстке регулювання журналістської діяльності, а також непропорційну відповідальність.

Також в Україні існує Концепція [43] розвитку штучного інтелекту, схвалена Кабінетом Міністрів, згідно з якою однією з пріоритетом застосування штучного інтелекту є інформаційна безпека. В ній ідеться про виявлення, запобігання та нейтралізацію наслідків недостовірної, неповної або упередженої інформації. Очікується, що штучний інтелект допоможе у виявленні потенційно небезпечної інформації, буде проводити аналіз інформації щодо авторства та джерела походження.

Також Рішенням [44] Ради національної безпеки і оборони України, введеного в дію Указом [45] президента, було створено Центр протидії дезінформації (ЦПД) як робочий орган РНБО. Згідно із Положенням [46] про ЦПД, затвердженим Указом президента, він буде, зокрема:

- проводити моніторинг та аналіз інфополя, виявляти ризики та небезпеки;
- брати участь у розробці стратегічної комунікації, надавати РНБО аналітичні дані, пропонувати концептуальні підходи щодо протидії дезінформації;
- брати участь у створенні системи оцінки інформаційних загроз та реагування на них;

• вивчати й узагальнювати практику та досвід інших країн, сприяти органам влади у протидії дезінформації, маніпулюванню громадською думкою тощо;

• виконувати інші завдання, передбачені положенням.

ЦПД не є виконавчим органом влади, не має повноважень із проведення перевірок, притягнення до відповідальності, а лише координуватиме роботу щодо протидії дезінформації і розроблятиме відповідні політики [47].

2.5. Огляд сучасних технологічних рішень для нейтралізації дезінформації

Найефективнішим способом розпізнавання фейкових новин та дезінформації є використання методів машинного навчання (MLT), які мають вирішальне значення для обробки величезної кількості інформації, що поширюється на платформах соціальних мереж. MLT – це підгалузь штучного інтелекту, яка спеціалізується на створенні алгоритмів та моделей, що дозволяють комп'ютерним системам навчатися та покращувати свою продуктивність завдяки досвіду, без необхідності явного програмування. Методи машинного навчання спеціально розроблені для аналізу та вилучення закономірностей з даних, що дозволяє системі виявляти та розуміти складні взаємозв'язки та робити обґрунтовані прогнози чи рішення. Різні алгоритми машинного навчання використовуються залежно від типу даних.

Головним чином, MLT охоплює наступне:

1) Методи глибокого навчання – це передові методи, що використовують штучні нейронні мережі з кількома шарами для отримання інформації та розпізнавання закономірностей у даних, виявлення та позначання фейкових новин шляхом аналізу лінгвістичних особливостей, присутніх у новинних статтях;

2) Метод обробки природної мови (NLP) – базується на алгоритмі, який дозволяє комп'ютерам ефективно обробляти, аналізувати та розуміти людську мову. Ці методи охоплюють широкий спектр завдань, включаючи класифікацію тексту, вилучення інформації, аналіз настроїв, машинний переклад та відповіді на

запитання. Робота з методами NLP вимагає використання статистичних моделей, алгоритмів машинного навчання та лінгвістичних ресурсів для вирішення складнощів та неоднозначностей, присутніх у природній мові;

3) Ансамблеве навчання – завдяки поєднанню різних моделей можна покращити загальну продуктивність, часто перевищуючи можливості будь-якої окремої моделі. Ансамблеве навчання довело свою ефективність у виявленні фейкових новин, комбінуючи результати різних алгоритмів, що призводить до підвищеної точності виявлення. Воно має досвід у вирішенні вразливостей у певних моделях та мінімізації будь-яких потенційних упереджень;

4) Трансферне навчання – дозволяє використовувати попередньо навчені моделі для нових завдань. Його використовували для виявлення фейкових новин шляхом застосування попередньо навчених моделей для завдань обробки природної мови та їх удосконалення спеціально для виявлення фейкових новин. Трансферне навчання – це цінний метод, який підвищує продуктивність учня, використовуючи знання з пов'язаної області. Досліджуючи реальні ситуації поза технічною сферою, можна отримати розуміння можливості трансферного навчання. Отримання навчальних даних, які точно відображають простір ознак та прогнозований розподіл тестових даних, може бути складним та дорогим у певних ситуаціях;

5) Графові методи – дозволяють аналізувати шаблони зв'язку, структури спільноти та властивості мережі. Графові алгоритми – це популярний метод представлення новинних статей у вигляді графіків та аналізу зв'язків між вузлами/вершинами, з'єднаними ребрами, з метою виявлення фейкових новин. Ці алгоритми продемонстрували обнадійливі результати у виявленні фейкових новин.

Загалом, машинне навчання є цінним інструментом для виявлення та припинення поширення неправдивої інформації в соціальних мережах.

Наприклад, надійні джерела новин використовують те саме доменне ім'я, що й джерело; тіньові використовують опечатки, щоб приховати це. У цьому випадку все

досить просто, оскільки навіть найпростіший запобіжний захід має вирішальне значення для захисту від загроз помилок. Щоб уникнути потрапляння на ці шахрайські сторінки, користувачі повинні бути уважними під час введення URL-адреси. Наприклад, країна походження (.ro, .de, .at, .uk) або мета (.eu, com, .edu) позначені закінченням URL-адреси та посиланнями в новинах. Також можна перевірити, кому належить цей домен, перейшовши за посиланням https://www.whois.net. Крім того, оскільки сфабриковані зображення часто підкріплюють неправдиву інформацію, можна рекомендувати дві альтернативи для розпізнавання порівнянних або візуально ідентичних веб-зображень: CLOUD Vision API, а саме алгоритм Google Images, та Tin Eye Reverse Image Search, сервіс, який надає рішення для пошуку зображень широкому колу користувачів.

Крім того, розширення браузера, відоме як Trusted Times, класифікує ненадійні та неправдиві новини. Воно надає додаткову інформацію про статті, включаючи упередженість та деталі про те, про кого висвітлюється позитивно та негативно, за допомогою машинного навчання. Воно здатне класифікувати статті як неправдиві, ненадійні, перевірені, або основні ЗМІ, узагальнення перевіреного контенту, визначення важливих тем, виявлення потенційної упередженості, визначення історичної упередженості репортерів та визначення історичної упередженості певного веб-сайту. Він застосовує значки для демонстрації надійності веб-сайтів.

Наприклад, виявлення клікбейт-заголовків може бути успішним шляхом створення класифікатора клікбейту на Python, використовуючи застосування процесів обробки даних OSEMN разом з алгоритмами NLP та машинного навчання.

Крім того, користувачам слід завантажувати файли або програмне забезпечення лише з офіційних веб-сайтів або перевірених джерел, щоб уникнути таких небезпек.

Ефективним, хоча й дорогим, методом захисту від помилок власників веб-сайтів є отримання торгових марок для їхніх доменів та придбання всіх пов’язаних

доменних імен, які схильні до помилок. Якщо домен зареєстровано як торгову марку, дійсно можна подати скарги на осіб, які займаються помилками [48].

Висновки до другого розділу 2

У другому розділі було розглянуто сучасні підходи та технології, що застосовуються для виявлення та нейтралізації дезінформації. Детально проаналізовано методи автоматичного виявлення фейкових новин, серед яких особливе місце займають моделі машинного навчання, у тому числі методи на основі штучного інтелекту, які здатні ефективно класифікувати інформаційні повідомлення за рівнем достовірності. Розглянуто також роль фактчекінгових платформ та алгоритмічних рішень, які забезпечують оперативну перевірку фактів у медіапросторі. Окрему увагу приділено державним та міжнародним ініціативам, які демонструють злагоджений підхід до боротьби з інформаційними загрозами на глобальному рівні. Огляд сучасних технологічних рішень, зокрема інтеграція штучного інтелекту, обробки природної мови та аналізу великих даних, показав перспективність таких інструментів для забезпечення інформаційної безпеки та підвищення стійкості суспільства до впливу дезінформації.

РОЗДІЛ 3. МОДЕЛЮВАННЯ ПОШИРЕННЯ ДЕЗІНФОРМАЦІЇ ТА СПОСОБИ ЇЇ НЕЙТРАЛІЗАЦІЇ

3.1. Основні механізми поширення дезінформації в цифровому середовищі

Дезінформація поширюється через складну взаємодію соціальних мереж, онлайн-сайтів новин, традиційних медіа та офлайн-просторів (Рисунок 9).

Наприклад, дезінформація може бути розміщена в інтернеті або висловлена публічною особою, поширюватися кількома дезінформаційними новинними сайтами (маскуючись під легітимні джерела новин), поширюватися через соціальні мережі та підхоплюватися традиційними новинними засобами, де вона може бути ненавмисно або цілеспрямовано посилена [49].

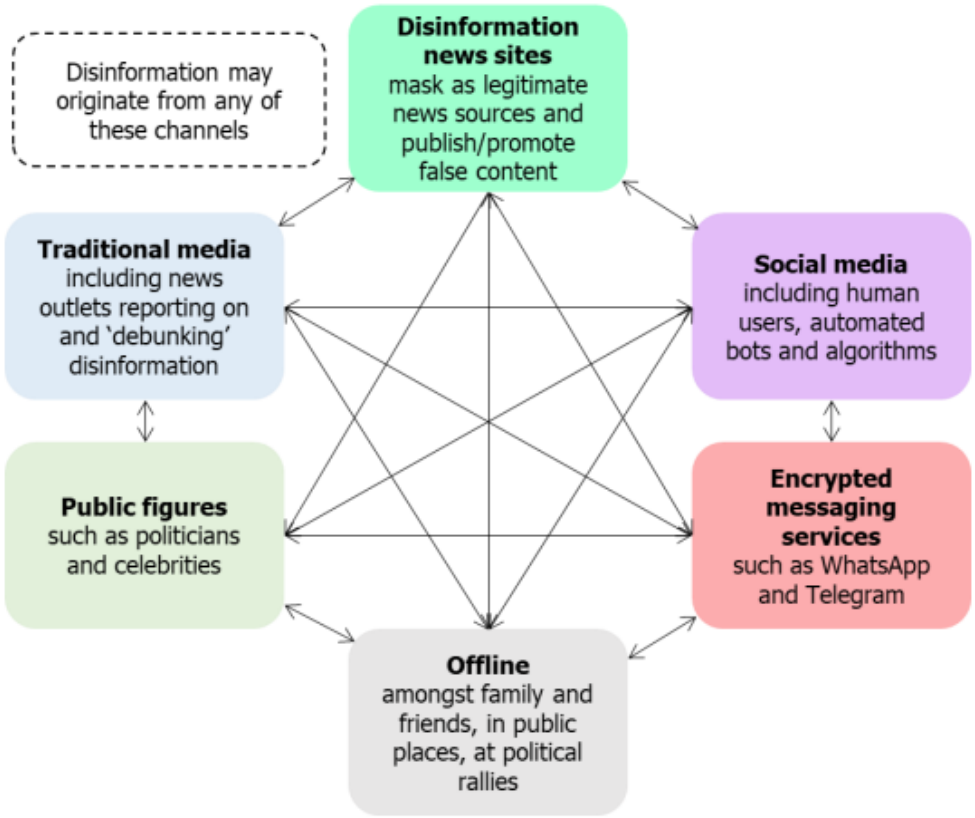

Рис. 9. Приклад того як дезінформація може поширюватись

Як ми бачимо вона поширюється як в цифровому, так і в фізичному середовищі, але в цьому розділі ми розглянемо тільки цифрове.

Академічні, урядові та галузеві зацікавлені сторони вважають зв'язок соціальних мереж та здатність будувати глобальні мережі центральними факторами поширення дезінформації/неправдивої інформації.

У листопаді 2023 року компанія з маркетингових досліджень Ipsos опитала користувачів Інтернету в 16 країнах і виявила, що 56% використовують соціальні мережі як основне джерело новин, а 68% кажуть, що дезінформація була найбільш поширеною саме там. Поширеність дезінформації відрізняється залежно від платформи [50].

Алгоритми соціальних мереж можуть збільшити поширення дезінформації, посилюючи контент з високою залученістю користувачів. «Боти» соціальних мереж, автоматизовані програми, які імітують поведінку людини в Інтернеті, також можуть використовуватися для лайків або кліків на дезінформацію у великих масштабах, щоб підвищити показники залученості («фермерство лайків/кліків»).

В одному дослідженні відстежувалося поширення перевірених фактами неправдивих новин у Twitter (тепер X) між 2006 і 2017 роками. У їхній вибірці неточні новини поширювалися швидше і ширше, ніж достовірна інформація в Інтернеті [51].

У 2023 році Ofcom визначив сервіси зашифрованих повідомлень, такі як WhatsApp, як ключовий канал поширення дезінформації. Сервіси, де користувачі можуть легко ділитися контентом далі, визначені як особливо ризиковані [52]. У 2020 році WhatsApp запровадив переадресацію міток після серії лінчів в Індії, пов'язаних з поширенням дезінформації в мережі [53].

Масштаби поширення важко кількісно оцінити, оскільки наскрізне шифрування означає, що доступ до контенту неможливий.

Зашифровані сервіси також недоступні для ініціатив з перевірки фактів. Закон про онлайн-безпеку 2023 року передбачає повноваження зобов'язувати сервіси

обміну повідомленнями перевіряти зашифрований контент. Однак такі платформи, як WhatsApp та Signal, висловлюють стурбованість тим, що це підриває конфіденційність користувачів [54].

Дезінформаційні новинні сайти можуть видавати себе за легітимні джерела новин та створювати або поширювати неправдивий та/або оманливий контент.

З лютого 2022 року NewsGuardi виявила 471 незаконний новинний сайт, що поширює дезінформацію, пов'язану з російсько-українською війною [55]. До березня 2024 року вони виявили 750 сайтів, створених штучним інтелектом, які поширюють неправдиві наративи та функціонують практично без людського контролю [56].

Наприклад, у листопаді 2023 року на новинному сайті, створеному штучним інтелектом, з'явилася неправдива стаття, в якій стверджувалося, що психіатр прем'єр-міністра Ізраїлю покінчив життя самогубством. Контент був взято з сатиричного блогу, опублікованого в 2010 році. Він транслювався по іранському державному телебаченню та поширювався на кількох платформах соціальних мереж, зокрема членом парламенту Індії [57].

Люди не обов'язково поширюють дезінформацію/неправдиву інформацію, бо вірять у неї.

У опитуванні 707 користувачів Facebook, проведеному у 2022 році, дослідники виявили, що 33% учасників поділилися неточними заголовками, бо вірили їм, 16% поділилися ними, незважаючи на те, що вони вважали їх неточними, а 51% поділилися ними через «неуважність до точності» [58].

Люди можуть свідомо поширювати неточний контент, щоб:

- просувати політичний порядок денний
- сигналізувати про спільну ідентичність
- розпалити дискусію чи дебати
- досягти соціального визнання
- викликати емоційну реакцію

• розважити інших (наприклад, гумором чи іронією)

• створити «хаос»

Опитування користувачів Facebook під час загальних виборів у США 2016 року показало, що 10% людей свідомо поширювали неправдиві/оманливі матеріали в Інтернеті [59], а 0,1% онлайн-користувачів відповідали за поширення 80% неправдивої/оманливої інформації [60].

Одне академічне дослідження показало, що більшість людей уникають навмисного поширення дезінформації, щоб захистити свою репутацію, навіть коли контент відповідає їхнім політичним поглядам [61]. Інше дослідження показало, що люди частіше діляться сумнівним контентом з друзями та родиною офлайн, ніж з великою аудиторією онлайн [62, 49].

Фейкові новини в Інтернеті також можуть поширюватися через ботів. Феррара та ін. (2016) у своєму розгляді соціальних ботів описують бота як «комп'ютерний алгоритм, який автоматично створює контент та взаємодіє з людьми в соціальних мережах, намагаючись наслідувати та, можливо, змінити їхню поведінку» [63].

Фальшиві новини можуть поширюватися через циклічні повідомлення, коли одне джерело публікує дезінформацію, яку підхоплює інше новинне видання, яке цитує оригінальне джерело як доказ того, що інформація є точною. Це продовжується, оскільки інші новинні агентства повідомляють про дезінформацію та увічнюють це замкнене коло [64].

У VAST Лі та Міллс Голуб відстежують контент понад 10 мільярдів веб-сайтів 75 мовами, щоб відстежувати, як контент поширюється в Інтернеті. Кампанії з дезінформації, за словами Лі та Міллс Голуба, поширюються чотирма ключовими способами:

Соціальна інженерія: забезпечення основи для неправильної характеристики та маніпулювання подіями, інцидентами, проблемами та публічним дискурсом. Соціальна інженерія часто спрямована на те, щоб схилити громадську думку на користь певного порядку денного.

Неавтентична ампліфікація: використання тролів, спам-ботів, фальшивих облікових записів, відомих як «шкарпеткові ляльки», платних облікових записів та сенсаційних лідерів думок для збільшення обсягу шкідливого контенту.

Мікротаргетинг: використання інструментів таргетування, розроблених для розміщення реклами та взаємодії з користувачами на платформах соціальних мереж, для виявлення та залучення найімовірніших аудиторій, які поширюватимуть та посилюватимуть дезінформацію.

Переслідування та зловживання: використання мобілізованої аудиторії, фейкових акаунтів та тролів для приховування, маргіналізації та придушення журналістів, протилежних поглядів та прозорого контенту [65].

3.2. Побудова простої моделі розповсюдження фейкових новин

Тепер ми спостерігаємо, що коли фейкові новини поширюються, наприклад, зловмисною особою, яка прагне ввести громадськість в оману, неправдива інформація накладається на іншу інформацію. Однак фейкові новини за своєю природою не представляють собою твердження правди про значення величини X, яке люди хочуть визначити, тому їх не можна розглядати як частину сигналу, який допомагає людям виявити справжнє значення X. З іншого боку, з точки зору обробки сигналів, все, що не є частиною сигналу, можна розглядати як шум. Дотримуючись цієї логіки, ми отримуємо нашу модель інформаційного процесу за наявності фейкових новин:

$$\eta t = \sigma\, X_t + B_t + F_t, \tag{1}$$

де часовий ряд $\{F_t\}$ представляє фейкові новини. Шумове визначення $\{B_t\}$, який не має зміщення, являє собою об'єднання великої кількості непідтверджених чуток та спекуляцій щодо значення X. Закон великих чисел тоді передбачає нормальність розподілу шуму, що робить броунівський рух життєздатним кандидатом для моделювання шуму. Таким чином, часовий ряд {Ft} вводить додаткове зміщення.

Визначення (фейкові новини). Часовий ряд $\{F_t\}$, що з'являється в інформаційному процесі (1), представляє «фейкові новини», якщо він має упередженість, таку як $E[F_t]\neq 0$, де E – математичне очікування.

Існування упередженості тут важливе, оскільки інакше це було б просто шумом, а не навмисною дезінформацією. Звичайно, додатковий неупереджений шум є неприємністю, затримуючи процес виявлення правди, але він не може зрештою відштовхнути громадськість від її виявлення. Що стосується статистичної залежності між $\{F_t\}$ та X, можуть виникнути дві ситуації: одна, коли ніхто не знає значення X, і в цьому випадку $\{F_t\}$ має бути незалежним від X, і одна, коли значення X відоме невеликій кількості осіб, які можуть бажати поширювати фейкові новини, і в цьому випадку $\{F_t\}$ цілком може залежати від X.

Ідея про те, що інформаційні моделі, представлені у вигляді $\xi t=\sigma X_t + B_t$ [66], можуть бути поширені на моделювання навмисних перекручень правди, розглядалася раніше [67]. Там пропонувалося, що зловмисник, який бажає маніпулювати громадськістю, може змінити значення швидкості потоку інформації σ. Отже, громадськість робитиме свої висновки на основі певного значення σ, тоді як фактичне значення σ насправді відрізняється, і як наслідок, громадськість вводиться в оману. Така схема зводиться до встановлення $F_t=\mu X_t$ для деякого μ, що може виникнути в моделі виборчої мікроструктури, описаній нижче, в якій значення X може бути відоме кандидату, але не громадськості, що дозволяє кандидату поширювати X-залежні фейкові новини. У більш загальному плані, враховуючи випадковість часу публікації, можна розглянути структуру фейкових новин виду $F_t=\mu X(t-\tau)1\{t > \tau\}$. Це еквівалентно наявності інформаційного процесу $\xi_t= \hat{}\sigma X_t + B_t$ з випадковим $\hat{}\sigma$, для якого можна отримати аналітичні вирази для умовних ймовірностей [67].

Для аналізу впливу фейкових новин буде корисно класифікувати членів громадськості на три різні категорії. Ми визначаємо Категорію I для позначення тих, хто не знає про потенційне існування фейкових елементів в інформації, яку

вони бачать. Тим не менш, вони діють раціонально, оскільки роблять свої оцінки відповідно до формули $P(X = x_i \mid \xi_t) = \frac{P(X=xi)\rho(\xi t \mid X=xi)}{\sum j\, P(X=xj)\rho(\xi t \mid X=xj)}$ [66], за винятком того, що η підставляється замість ξt. Іншими словами, вони «правильно» роблять висновок про апостеріорну ймовірність, але на основі помилкового переконання, що інформація, яку вони отримують, має тип $\xi_t = \sigma X_t + B_t$, тоді як насправді вона має тип (1). Як ми побачимо, люди цієї категорії найбільш вразливі до впливу фейкових новин. Ми позначаємо Категорією II тих представників громадськості, які знають про потенційне існування фейкових новин, але не знають точного часу, коли елементи фейкових новин у часовому ряді $\{F_t\}$ публікуються. Ці особи стикаються з найскладнішим технічним завданням, оскільки в їхній оцінці вони повинні мати справу з трьома невідомими, X, $\{B_t\}$ та $\{F_t\}$, але лише з однією відомою, $\{\eta_t\}$. Як ми побачимо нижче, хоча аналітичні вирази для умовної ймовірності $P(X=x_i|\{\eta_s\}0\leq s \leq t)$ можна отримати, аналіз є більш складним, ніж для виборців Категорії I. Таким чином, люди цієї категорії значно більше усвідомлюють невизначеності у своїх оцінках порівняно з тими, хто належить до Категорії I. Нарешті, Категорія III складається з тих людей, які є високопоінформованими настільки, що знають значення часового ряду $\{F_t\}$. Оскільки $\{F_t\}$ не містить інформації, що стосується X, вони можуть просто не враховувати $\{F_t\}$ зі своєї інформації $\{\eta_t\}$ та використовувати $\xi_t = \eta_t - F_t$ замість цього, щоб розрахувати своє апостеріорне переконання відповідно до $P(X = x_i \mid |\xi_t) = \frac{pi\ exp\ (\sigma xi\xi t - \frac{1}{2}\sigma^2 xi^2 t)}{p0 + p1\ exp\ (\sigma\xi t - \frac{1}{2}\sigma^2 t)}$ [66]. Як і ті, хто належить до Категорії I, люди Категорії III, ймовірно, будуть наполегливими щодо своїх суджень. Однак зазначимо, що особу Категорії III слід певною мірою розглядати як ідеалізацію. Зрештою, для будь-якої особи майже нездоланним завданням є ідеальне визначення того, які новини є фальшивими, а які ні [66].

3.3. Методи виявлення та аналізу дезінформації

Перш ніж почати розуміти, як ідентифікувати фейкові новини, дослідник повинен виробити всебічне розуміння їх різноманіття. Крім того, важливо вивчити причини та мотивацію поширення фейкових новин, такі як зловмисники, фінансові стимули та алгоритми соціальних мереж. Більше того, важливо знати про наслідки фейкових новин, такі як підрив довіри до ЗМІ, складність політичного дискурсу та потенціал для підбурювання до насильства. Щойно дослідник отримає глибоке розуміння фейкових новин, він зможе почати досліджувати способи їх виявлення. Різні дослідники мають різні погляди на вивчення фейкових новин. Вивчення фейкових новин може мати різні форми. Багато дослідників досліджують відсоток ясності в новинах, тоді як інші досліджують, як створюються фейкові новини. І навпаки, багато досліджень безпосередньо досліджують шляхи поширення фейкових новин, тоді як інші досліджують достовірність джерела. Існує чотири основні перспективи дослідження виявлення фейкових новин, як показано на рисунку 1. У двох перспективах фейкові новини вивчаються під час їх створення: дослідження на основі знань та стилю, а в двох перспективах – після їх поширення: поширення та дослідження на основі джерел.

A. ДОСЛІДЖЕННЯ НА ОСНОВІ ЗНАНЬ

Підходи, засновані на знаннях, також відомі як перевірка фактів, передбачають збір необроблених фактів з відкритої мережі для побудови знань. Однак ці дані необхідно додатково обробляти для вирішення проблем надмірності, недійсності, конфліктів, ненадійності та неповноти. Для забезпечення точності фактів необхідні п'ять тестів: роздільна здатність сутностей, реєстрація часу, оцінка узгодженості, оцінка повноти та оцінка точності. Роздільна здатність сутностей використовується для зв'язування записів подібних фактів, тоді як реєстрація часу використовується для вимірювання часової достовірності фактів. Оцінка узгодженості гарантує, що факти не суперечать одна одній, оцінка повноти гарантує, що всі факти враховані, а оцінка точності гарантує, що факти є правильними. Дані можна ефективно

очистити, виконавши ці п'ять тестів, і успішно впровадити підходи, засновані на знаннях.

Б. ДОСЛІДЖЕННЯ НА ОСНОВІ СТИЛЮ

Мета досліджень на основі стилю полягає в оцінці, чи призначені новини для введення громадськості в оману. Підхід, заснований на стилі, намагається охопити стиль написання новинного контенту. Можна відрізнити фейкові новини від правди, визначаючи стилі на основі деяких кількісних характеристик. Фейкові новини ідентифікуються за допомогою теорій, шаблонів та стратегій. Текст і зображення відрізняють фейкові новини від справжніх новин. Текстові шаблони стосуються неформальності, різноманітності, суб'єктивності та натхненної мови. У фейкових зображеннях ясність та узгодженість часто високі, але різноманітність часто низька через відсутність балів кластеризації.

В. ДОСЛІДЖЕННЯ НА ОСНОВІ ПОШИРЕННЯ

Дослідження на основі поширення намагаються зрозуміти, як правдива інформація поширюється в мережах, де новини, створені аномаліями, прогнозуються як неправдива інформація. Цей вид дослідження головним чином стосується того, як користувач бере участь у поширенні фейкових новин. Вхідними даними такого дослідження можуть бути або каскад новин, або самовизначений граф. У каскаді новин поширення новин представлено безпосередньо, тоді як самовизначені графи є непрямими представленнями, які фіксують більше інформації про поширення. Каскад – це структура, подібна до дерева, яка визначається ініціатором фейкових новин. Він складається з кількох вузлів, кожен з яких визначається користувачами, які опублікували або переслали фейкові новини. І навпаки, самовизначений граф – це мережа, в якій поширюються фейкові новини, і може бути різних типів. Однорідна мережа – це один тип мережі, що складається з одного типу вузла та ребра. Гетерогенна мережа – це інший тип, що складається з кількох вузлів та ребер, і зазвичай є складнішою, ніж однорідна

мережа. Нарешті, ієрархічна мережа складається з вузлів, систематично розташованих у шарах або ярусах, де кореневий вузол ініціює фейкові новини.

Г. ДОСЛІДЖЕННЯ НА ОСНОВІ ДЖЕРЕЛ

Дослідження на основі джерел досліджує достовірність його творців та розповсюджувачів. Дослідження може бути проведене на основі новин та соціальної інформації. Достовірність джерела можна оцінити, спостерігаючи за роллю автора новин та видавця. Зловмисні користувачі соціальних мереж можуть ініціювати та поширювати новини всередині та за межами своїх мереж. Звичайні користувачі пересилають або поширюють сфабриковані новини без будь-якої перевірки. Дослідження на основі джерел стосуються того, як користувачі взаємодіють з фейковими новинами та їхньої ролі в їх створенні, публікації та поширенні в ЗМІ. Існує кілька типів досліджень, які вивчають життєвий цикл фейкових новин, природу їх поширення, причини поширення та проблеми виявлення, а також досліджують різні перспективи досліджень фейкових новин, як описано в таблиці 1. Наприклад, за даними Бонд'єллі та ін., фейкові новини можна визначити багатьма способами, зібрати багатьма способами та засвідчити за допомогою різних методів. Вони повідомили, що існує обмежений набір даних, який можна розглядати як стандартний у будь-який необхідний спосіб. Вони заявили, що методи глибокого навчання дуже ефективні у виявленні фейкових новин [68].

3.4. Підходи до обмеження поширення неправдивої інформації

1. Уникайте повторення дезінформації без виправлення

Повторення неправдивих тверджень посилює віру в ці твердження, явище, відоме як ефект ілюзорної правди. Люди будь-якого віку схильні до ілюзорної правди, навіть якщо вони вже мають відповідні попередні знання з цієї теми. Коли медіа-джерела, політичні еліти або знаменитості повторюють дезінформацію, їхній

вплив та повторення можуть увічнювати хибні переконання. Повторення дезінформації необхідне лише тоді, коли активно виправляється неправда. У цих випадках неправду слід повторювати коротко, при цьому виправлення має бути помітнішим, ніж сама неправда.

2. Співпрацюйте з компаніями соціальних мереж, щоб зрозуміти та зменшити поширення шкідливої дезінформації

Більшість дезінформації в соціальних мережах поширюється дуже малою кількістю користувачів, навіть під час надзвичайних ситуацій у сфері охорони здоров'я. Ці «суперпоширювачі» можуть відігравати надзвичайно велику роль у поширенні дезінформації. «Ехо-камери» соціальних мереж зв'язують та ізолюють спільноти зі схожими переконаннями, що сприяє поширенню неправди та перешкоджає поширенню фактичних виправлень. У соціальних мережах сенсаційний, моральний, емоційний та зневажливий контент про «іншу сторону» може поширюватися швидше, ніж нейтральний або позитивний контент. Вчені, політики та фахівці у сфері охорони здоров'я повинні співпрацювати з онлайн-платформами, щоб зрозуміти та використати структури стимулів соціальних мереж для зменшення поширення небезпечної дезінформації.

3. Використовуйте стратегії виправлення дезінформації з інструментами, які вже довели свою ефективність у сприянні здоровій поведінці.

Дослідження психологічних наук показують, що зв'язок між знаннями та поведінкою є недосконалим. Існують вагомі докази того, що обмеження хибних уявлень може змінити основні переконання та ставлення, пов'язані зі здоров'ям, але цього може бути недостатньо для зміни поведінки та прийняття рішень у реальному світі. Виправлення дезінформації за допомогою точних рекомендацій щодо здоров'я є життєво важливим, але воно має відбуватися разом зі стратегіями, заснованими на доказах, які сприяють здоровій поведінці (наприклад, консультування, навчання навичкам, стимули, соціальні норми).

4. Використовуйте довірені джерела для протидії дезінформації та надання точної інформації про здоров'я.

Люди вірять у дезінформацію та поширюють її з багатьох причин: вони можуть вважати її відповідною своїй соціальній чи політичній ідентичності, вони можуть не враховувати її точність, або вони можуть вважати її цікавою чи корисною. Ці мотивації є складними та часто взаємопов'язаними. Спроби виправити дезінформацію та зменшити її поширення є найбільш успішними, коли інформація надходить з довірених джерел та представників, включаючи релігійних, політичних та громадських лідерів.

5. Часто та неодноразово спростовуйте дезінформацію, використовуючи методи, засновані на доказах

Дослідження показують, що спростування дезінформації, як правило, ефективне для різних вікових груп та культур. Однак спростування не завжди повністю усуває хибні уявлення. Виправлення повинні бути помітними разом із дезінформацією, щоб точна інформація правильно зберігалася та витягувалася з пам'яті. Спростування є найефективнішим, коли воно походить з надійних джерел, надає достатньо деталей про те, чому твердження є хибним, та пропонує вказівки щодо того, що є правдою. Оскільки ефективність спростування з часом зменшується, його слід повторювати через надійні канали та методи, засновані на доказах.

6. Попереднє спростування дезінформації для щеплення вразливої аудиторії шляхом розвитку навичок та стійкості з раннього віку

Замість того, щоб виправляти дезінформацію після факту, «попереднє спростування» має бути першою лінією захисту для формування стійкості громадськості до дезінформації заздалегідь. Дослідження показують, що психологічні втручання з щеплення можуть допомогти людям виявити окремі приклади дезінформації або загальні методи, які зазвичай використовуються в дезінформаційних кампаніях. Попереднє охоплення аудиторії може бути

масштабоване, щоб охопити мільйони людей у соціальних мережах за допомогою коротких відео чи повідомлень, або ж його можна проводити у формі інтерактивних інструментів, що включають ігри чи вікторини. Однак, з часом ефект від попереднього охоплення аудиторії згасає; для підтримки стійкості до дезінформації можуть знадобитися регулярні «підсилювачі», а також навчання медіа та цифровій грамотності.

7. Вимагайте від компаній соціальних мереж доступу до даних та прозорості для проведення наукових досліджень дезінформації

Зусилля щодо кількісної оцінки та розуміння дезінформації в соціальних мережах ускладнюються відсутністю доступу до даних користувачів від компаній соціальних мереж. Втручання у боротьбу з дезінформацією рідко тестуються в реальних умовах через аналогічну відсутність співпраці з галуззю. Загальнодоступні дані пропонують обмежений знімок впливу, але вони не можуть пояснити вплив на населення та мережу. Дослідникам потрібен доступ до повного переліку публікацій у соціальних мережах на різних платформах, а також дані, що розкривають, як алгоритми формують те, що бачать окремі користувачі. Відповідальний обмін даними може використовувати рамки, що використовуються зараз, для управління конфіденційними медичними даними. Політики та органи охорони здоров'я повинні заохочувати дослідницькі партнерства та вимагати від компаній соціальних мереж більшого контролю та прозорості щоб стримати поширення дезінформації.

8. Фінансувати фундаментальні та трансляційні дослідження в галузі психології дезінформації про здоров'я, включаючи ефективні способи протидії їй

Було розроблено кілька втручань для протидії дезінформації про здоров'я, але дослідники ще не порівняли їхні результати, окремо чи в поєднанні. Існує потреба зрозуміти, які втручання ефективні для конкретних типів інформації: те, що працює для однієї проблеми, може не застосовуватися до інших. В ідеалі, на ці питання слід відповісти за допомогою масштабних випробувань з репрезентативною цільовою

аудиторією в реальних умовах. Для вирішення цих важливих питань щодо цифрового життя необхідні додаткові можливості фінансування досліджень у галузі психологічних наук [69].

3.5. Приклади застосування методів нейтралізації дезінформації

Facebook, який зазнав жорсткої критики за те, що дозволив поширювати фейкові новини під час виборчого періоду, вжив заходів для боротьби з цією проблемою.

Одним із таких кроків є залучення Міжнародної мережі перевірки фактів (IFCN), філії флоридського аналітичного центру журналістики Poynter, зовнішнє джерело. Користувачі Facebook у США та Німеччині тепер можуть позначати статті, які вони вважають навмисно неправдивими, а потім вони потраплятимуть до сторонніх перевіряльників фактів, зареєстрованих в IFCN.

Ці перевіряльники фактів надходять від медіа-організацій, таких як Washington Post, та веб-сайтів, таких як сайт розвінчування міських легенд Snopes.com, зовнішнє джерело.

Сторонні перевіряльники фактів, каже директор IFCN Алексіос Манцарліс, «розглядають історії, які користувачі позначили як фейкові, і якщо вони перевіряють їх та позначають як неправдиві, ці історії отримують спірний тег, який залишається з ними в соціальній мережі».

Ще одне попередження з'являється, якщо користувачі намагаються поділитися історією, хоча Facebook не запобігає такому поширенню та не видаляє фейкову новину. Однак тег «фейк» негативно вплине на рейтинг історії в алгоритмі Facebook, а це означає, що менше людей побачать її у своїх стрічках новин.

Мантцарліс каже, що поки що немає вагомих доказів того, що це насправді зупиняє поширення фейкових новин у великих масштабах, і є питання щодо того, наскільки стійкою може бути ця програма. Facebook не платить членам IFCN за

надання послуг з перевірки фактів. Також ніщо не заважає фейковим новинам публікуватися та поширюватися взагалі – і, можливо, досить широко, перш ніж їх позначать тегом.

«Існує затримка, тому доки історію не позначать як неправдиву, вона продовжує поширюватися в соціальній мережі», – каже Манцарліс.

У 2009 році Le Monde, одна з найбільших французьких газет, створила підрозділ перевірки фактів під назвою Les Decodeurs, external (Декодери). Цей підрозділ все частіше звертає свою увагу на фейкові новини та розробив веб-розширення під назвою Decodex [70].

«Ви просто встановлюєте його у своєму браузері, і коли переходите на сайт фейкових новин, з'являється спливаюче вікно з написом «попередження: це сайт фейкових новин», — каже Самюель Лоран, редактор Les Decodeurs. «Якщо ви натиснете на інструмент, ви отримаєте доступ до невеликої статті з описом веб-сайту та поясненням, чому ми вважаємо його ненадійним».

Розширення пов'язане з базою даних, складеною Les Decodeurs, яка ранжує сайти як «фейкові», «справжні» або «сатиричні».

Але існує кілька перешкод, які, ймовірно, заважають відносно простому програмному забезпеченню стати чарівною паличкою для фейкових новин. Користувачі, по-перше, повинні знати про проблему неправдивих історій. Вони повинні знати про розширення та бути достатньо стурбованими, щоб завантажити його. І вони повинні довіряти журналістам Le Monde та лівоцентристській перспективі газети.

«Ми знаємо, що не переконаємо всіх, і ми знаємо, що читачі фейкових новин вже вважають нас фейковими новинами», — каже Лоран. «Наша мета — просто надати цей інструмент людям, які справді скептично налаштовані або просто не знають, кому довіряти.

«Людям, які вже потрапили у вир фейкових новин, вже надто пізно» [71].

Висновки до третього розділу

У третьому розділі було проведено дослідження механізмів поширення дезінформації в цифровому середовищі та методів її нейтралізації. На основі аналізу встановила, що основними чинниками поширення фейкових новин є використання соціальних платформ, алгоритмічне підсилення контенту, автоматизовані боти та слабка цифрова грамотність користувачів.

Була збудована спрощена модель розповсюдження фейкових новин дозволила імітувати динаміку поширення неправдивого контенту та продемонструвала, як швидко такі повідомлення можуть охоплювати великі масиви аудиторії. Це підкреслює необхідність оперативного виявлення джерел дезінформації.

У розділі також розглянула сучасні підходи до виявлення дезінформації — зокрема, використання алгоритмів машинного навчання, фактчекінгу та систем оцінювання достовірності джерел. Це показало, що ефективність таких методів значно зростає за умови поєднання автоматизованих інструментів з експертною оцінкою.

Окрему увагу в цьому розділі було приділено підходам до обмеження поширення неправдивої інформації. Розглянула психологічно обґрунтовані методи, засновані на результатах досліджень. Зокрема, акцентувала на необхідності уникати повторення дезінформації без чіткого виправлення, важливості використання достовірних джерел, попередньому інформуванні аудиторії, а також на потребі співпраці з онлайн-платформами для підвищення прозорості алгоритмів.

Отримані результати підтверджують, що боротьба з дезінформацією має бути комплексною, поєднувати технологічні рішення, освітні ініціативи та міжсекторальну співпрацю для забезпечення стійкості до інформаційних загроз.

ВИСНОВКИ

У дипломній роботі було всебічно розглянуто процеси поширення дезінформації в цифровому середовищі, а також методи її виявлення та нейтралізації. У першому розділі проаналізовано теоретичні засади феномену дезінформації, зокрема її поняття, класифікацію та основні види. Встановлено, що дезінформація є складним соціально-комунікативним явищем, яке активно використовується як інструмент впливу на суспільну думку. Окрему увагу приділено механізмам поширення фейкових новин та ролі соціальних мереж, ботів і фейкових акаунтів у поширенні неправдивого контенту.

У другому розділі проведено аналіз сучасних підходів до протидії дезінформації. Розглянуто методи автоматичного виявлення фейкових новин, використання алгоритмів штучного інтелекту, фактчекінг і роль міжнародних ініціатив у стримуванні інформаційних загроз. Показано, що ефективна боротьба з дезінформацією потребує поєднання технологічних рішень, нормативно-правових заходів, а також просвітницької роботи з підвищення рівня цифрової грамотності населення.

У третьому розділі було змодельовано процес поширення фейкових новин, проаналізовано ключові механізми їх циркуляції в цифровому середовищі та розглянуто практичні методи їх нейтралізації. Визначено, що ефективна протидія дезінформації можлива за умов своєчасного виявлення джерел поширення, залучення перевірених комунікаційних каналів, а також впровадження стратегій попереднього інформування та спростування.

Загалом, результати дослідження підтверджують, що дезінформація є серйозною загрозою для інформаційної безпеки, яка потребує системного підходу до виявлення, аналізу та протидії. Практична цінність роботи полягає у поєднанні теоретичних моделей поширення з аналізом реальних механізмів нейтралізації, що

може бути використано в подальших дослідженнях або прикладних розробках у сфері кібербезпеки та інформаційної гігієни.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Беднарчук О. Вплив дезінформації на суспільство: правові аспекти протидії. Право і суспільство. 2023. № 4. С. 762–768. URL: https://applaw.net/index.php/journal/article/download/762/668/

2. Protocols of the Elders of Zion // Holocaust Encyclopedia. Washington: United States Holocaust Memorial Museum, 2023. URL: https://encyclopedia.ushmm.org/content/uk/article/protocols-of-the-elders-of-zion

3. Shapiro J. Disinformation and democracy: a historical overview. Journal of Democracy. 2015. Vol. 26, № 2. P. 5–16. URL: https://muse.jhu.edu/article/579342

4. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., et al. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining. С. 849–857. URL: https://cse.buffalo.edu/~lusu/papers/KDD2018Yaqing.pdf

5. Fetzer, J. H. Disinformation: The use of false information. Minds and Machines. 2004. С. 231–240. Springer. URL: https://www.researchgate.net/publication/225841588_Disinformation_The_Use_of_False_Information

6. Hernon, P. Disinformation and misinformation through the internet: Findings of an exploratory study. Government Information Quarterly, 1995. С. 133–139. Elsevier BV. URL: https://doi.org/10.1016%2F0740-624x%2895%2990052-7

7. Allcott, H., & Gentzkow, M. Social Media and Fake News in the 2016 Election. National Bureau of Economic Research. 2017. URL: https://doi.org/10.3386%2Fw23089

8. Pratiwi, I. Y. R., Asmara, R. A., & Rahutomo, F. Study of hoax news detection using na¨ıve bayes classifier in Indonesian language. In 2017 11th International Conference on Information & Communication Technology and System (ICTS). 2017.

URL: https://www.researchgate.net/publication/322649299_Study_of_hoax_news_detection_using_naive_bayes_classifier_in_Indonesian_language

9. Rosnow, R. L. Inside rumor: A personal journey.. American Psychologist, C. 484-496. American Psychological Association (APA). 1991. URL: https://doi.org/10.1037%2F0003-066x.46.5.484

10. Silverman C. Analysis of disinformation in social media: machine learning methods // ArXiv. 2020. URL: https://arxiv.org/pdf/2007.07388

11. Disinformation and 7 common forms of information disorder. London: Commons Library, 2023. URL: https://commonslibrary.org/disinformation-and-7-common-forms-of-information-disorder/

12. Social media users. DataReportal, 2025. URL: https://datareportal.com/social-media-users

13. Fake news & cyber propaganda: the abuse of social media. Trend Micro, 2022. URL: https://www.trendmicro.com/vinfo/de/security/news/cybercrime-and-digital-threats/fake-news-cyber-propaganda-the-abuse-of-social-media

14. Механізми протидії дезінформації. Київ: Центр протидії дезінформації, 2024. URL: https://cpd.gov.ua/category/glossary/mechanisms/

15. Wardle C. Strategies for combating disinformation in public health // Health Promotion International. 2025. Vol. 40, № 2. URL: https://academic.oup.com/heapro/article/40/2/daaf023/8100645

16. Background paper on disinformation and fake news. Geneva: Parliamentary Assembly on Migration, 2024. URL: https://pam.int/wp-content/uploads/2024/10/EN-Background-paper-on-disinformation-and-fake-news-Jan-2024.pdf

17. Johnson E. How technology fuels the spread of fake news. Financial Times. 2023. URL: https://www.ft.com/content/0aa9725d-e423-4a6b-b842-866ad4541dc2

18. Ковальчук А. Моделі виявлення фейкових новин у соціальних мережа х. Вісник ВНТУ. 2023. № 1. С. 1-10.
URL: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/41427/150136.pdf?sequenc e=2&isAllowed=y

19. What are fake news? London: BBC Newsround, 2024. URL: https://www.b bc.co.uk/newsround/69009887

20. Breaking news: how disinformation reshapes media. Radiolab. 2023. URL: https://radiolab.org/podcast/breaking-news/transcript

21. Disinformation in the digital age: committee report. London: UK Parliamen t, 2023. URL: https://committees.parliament.uk/writtenevidence/136915/pdf/

22. Review of methods for combating disinformation. Open Research Knowledge Graph. 2024. URL: https://orkg.org/review/R756064

23. Brown T. Artificial intelligence technologies for detecting disinformation. Data in Brief. 2022. Vol. 38. URL: https://www.sciencedirect.com/scien ce/article/pii/S2405959521001375#b15

24. Smith J. Using sensors to monitor disinformation. Sensors. 2024. Vol. 24, № 11. URL: https://www.mdpi.com/1424-8220/24/11/3590

25. Gupta R. A machine learning technique for detection of social media fake n ews .ResearchGate. 2023. URL: https://www.researchgate.net/publication/372547193_A _Machine_Learning_Technique_for_Detection_of_Social_Media_Fake_News

26. Naive Bayes classifier for fake news detection. GitHub. 2023. URL: https:// github.com/bhroben/Naive-Bayes-multinomial-classifier-for-fake-news-detection/blob/main/Naive_Bayes_classifier_for_Fake_News.ipynb

27. Іванов О. Виявлення дезінформації за допомогою біоінформатики.BIO Web of Conferences. 2024. URL: https://www.bio-conferences.org/articles/bioconf/pdf/2024/16/bioconf_iscku2024_00049.pdf

28. Brown J. Logistic regression for machine learning. Machine Learning Mastery. 2023. URL: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

29. Kuntala C. Fake news detection using logistic regression: ML project tutorial for beginners.Medium. 2023. URL: https://medium.com/@chandanakuntala21/fake-news-detection-using-logistic-regression-ml-project-tutorial-beginners-89b5c704d6ee

30. Lee S. Real-time disinformation detection algorithms.IEEE Xplore. 2023. URL: https://ieeexplore.ieee.org/document/10182240

31. Rashidi S. How to use AI for data analysis: a step-by-step guide.Forbes. 2024. URL: https://www.forbes.com/sites/solrashidi/2024/10/16/how-to-use-ai-for-data-analysis-a-step-by-step-guide/

32. Quick guide to spotting misinformation. New York: UNICEF, 2023. URL: https://www.unicef.org/eca/stories/quick-guide-spotting-misinformation

33. Countering disinformation: Council of Europe resolution. Strasbourg: Council of Europe, 2021. URL: https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-EN.asp?fileid=28598&lang=en

34. Code of practice on disinformation. Brussels: European Commission, 2022. URL: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

35. Законодавство України про протидію дезінформації. Київ: Верховна Рада України, 2023. URL: https://thedigital.gov.ua/storage/uploads/files/normative_document/archive/5bcf04496e848695859521.pdf

36. Finland fights fake news: interactive report.CNN. 2019. URL: https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

37. Portail officiel du gouvernement français. Paris: Gouvernement de France, 2025. URL: https://www.info.gouv.fr/

38. Facebook baseline report on implementation of the code of practice on disinformation. Brussels: European Commission, 2019. URL: https://ec.europa.eu/information_society/newsroom/image/document/2019-5/facebook_baseline_report_on_implementation_of_the_code_of_practice_on_disinformation_CF161D11-9A54-3E27-65D58168CAC40050_56991.pdf

39. Про інформацію: Закон України від 02.10.1992 № 2657-XII. URL: https://zakon.rada.gov.ua/laws/show/2657-12/conv#n24

40. Про друковані засоби масової інформації (пресу) в Україні: Закон України від 16.11.1992 № 2782-XII. URL: https://zakon.rada.gov.ua/laws/show/2782-12/conv#n186

41. Про рекламу: Закон України від 03.07.1996 № 3759-XII. URL: https://zakon.rada.gov.ua/laws/show/3759-12/conv#n908

42. Офіційний сайт Міністерства культури та інформаційної політики України. Київ: МКІП, 2025. URL: https://mcsc.gov.ua/

43. Про схвалення Концепції розвитку штучного інтелекту в Україні: Розпорядження КМУ від 02.12.2020 № 1556-р. URL: https://www.kmu.gov.ua/npas/pro-shvalennya-koncepciyi-rozvitku-shtuchnogo-intelektu-v-ukrayini-s21220

44. Про затвердження Порядку формування та використання електронних інформаційних ресурсів: Наказ МКІП від 15.03.2021 № 155. URL: https://zakon.rada.gov.ua/laws/show/n0015525-21#Text

45. Про заходи щодо протидії дезінформації: Указ Президента України від 15.03.2021 № 106/2021. URL: https://zakon.rada.gov.ua/laws/show/106/2021#n5

46. Про Стратегію інформаційної безпеки: Указ Президента України від 15.05.2021 № 187/2021. URL: https://zakon.rada.gov.ua/laws/show/187/2021#n15

47. Епідемія дезінформації: аналітичний звіт. Київ: ЦЕДЕМ, 2023. URL: https://cedem.org.ua/analytics/epidemiya-dezinformatsiyi/

48. Petrov O. Digital technologies used to combat disinformation and fake new s.ResearchGate. 2024. URL: https://www.researchgate.net/publication/380987988_Digital_Technologies_Used_to_Combat_Disinformation_and_Fake_News

49. Disinformation: a short overview. London: Parliamentary Office of Science and Technology, 2023. URL: https://researchbriefings.files.parliament.uk/documents/POST-PN-0719/POST-PN-0719.pdf

50. Ipsos. 2023. Elections, social media: The battle against disinformation and trust issues. URL: https://www.ipsos.com/en/elections-social-media-battle-against-disinformation-and-trust-issueshttps://www.science.org/doi/10.1126/science.aap9559

51. Vosoughi, S. The spread of true and false news online// Vosoughi, S., Roy, D., & Aral, S./ — 2018. — URL: 10.1126/science.aap9559

52. How misinformation on WhatsApp led to a mob lynching in India / The Washington Post. — 2020. — URL: https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india/

53. The UK just passed an online safety law that could make people less safe / The Conversation. — 2023. — URL: https://theconversation.com/the-uk-just-passed-an-online-safety-law-that-could-make-people-less-safe-213595

54. Russian Disinformation Tracking Center / NewsGuard. — URL: https://www.newsguardtech.com/special-reports/russian-disinformation-tracking-center/

55. AI Tracking Center / NewsGuard. — URL: https://www.newsguardtech.com/special-reports/ai-tracking-center/

56. AI-generated site sparks viral hoax claiming the suicide of Netanyahu’s purported psychiatrist / NewsGuard. — URL: https://www.newsguardtech.com/special-reports/ai-generated-site-sparks-viral-hoax-claiming-the-suicide-of-netanyahus-purported-psychiatrist/

57. Altmann, D. M. The immunology of misinformation: Why fake news about vaccines goes viral / D. M. Altmann, R. J. Boyton // Trends in Immunology. — 2021. — URL: https://www.sciencedirect.com/science/article/pii/S1364661321000516

58. Lewandowsky, S. Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making / S. Lewandowsky, L. Smillie // Psychological Medicine. — 2019. — URL: https://pubmed.ncbi.nlm.nih.gov/30662946/

59. Vosoughi, S. The spread of true and false news online / S. Vosoughi, D. Roy, S. Aral // Science. — 2018. — URL: https://www.science.org/doi/10.1126/science.aau2706

60. Pennycook, G. Why do so few people share fake news? It hurts their reputation / G. Pennycook, D. G. Rand // ResearchGate. — 2020. — URL: https://www.researchgate.net/publication/346311907_Why_do_so_few_people_share_fake_news_It_hurts_their_reputation

61. Offline vs. online sharing: How modes of communication affect the spread of misinformation / Centre for the Study of African Politics. — URL: https://www.csap.cam.ac.uk/media/uploads/files/1/offline-vs-online-sharing.pdf

62. Ferrara, E. The rise of social bots / E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini // ResearchGate. — 2014. — URL: https://www.researchgate.net/publication/264123205_The_Rise_of_Social_Bots

63. How fake news spreads. Victoria: University of Victoria, 2023. URL: https: //libguides.uvic.ca/fakenews/how-it-spreads

64. Four ways disinformation campaigns are propagated online.World Economi c Forum. 2022. URL: https://www.weforum.org/stories/2022/08/four-ways-disinformation-campaigns-are-propagated-online/

65. Wilson T. How to model fake news.ResearchGate. 2018. URL: https://www .researchgate.net/publication/327435057_How_to_model_fake_news

66. Brody, D. C. Asset pricing with random information flow / D. C. Brody, L. P. Hughston, A. Macrina // ResearchGate. — 2018. — URL: https://www.researchgate.net/profile/Dorje-Brody/publication/46586119_Asset_pricing_with_random_information_flow/links/5b052c350f7e9b24a2af6eb6/Asset-pricing-with-random-information-flow.pdf

67. Singh A. A review of methodologies for fake news analysis.ResearchGate. 2023. URL: https://www.researchgate.net/publication/372363091_A_review_of_methodologies_for_fake_news_analysis

68. Misinformation recommendations. Washington: American Psychological Association, 2023. URL: https://www.apa.org/topics/journalism-facts/misinformation-recommendations

69. Les Décodeurs / Le Monde. — URL: https://www.lemonde.fr/les-decodeurs/

70. How fake news goes viral.BBC. 2017. URL: https://www.bbc.com/news/blogs-trending-38769996