УДК 004.056.5:004.032.26:070.16 DOI: 10.31673/2412-4338.2024.024860

Рабчун Дмитро Ігорович
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID: 0000-0002-5555-0910

Тищенко Віталій Сергійович
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID ID: 0000-0003-3849-6243

Голобородько Сергій Олександрович
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID ID:0009-0006-9892-3031

ЕФЕКТИВНЕ РОЗПІЗНАВАННЯ ДЕЗІНФОРМАЦІЇ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ: ФОКУС НА ВИЯВЛЕННІ ЕМОЦІЙНОГО ВПЛИВУ

Анотація. *Дослідження фокусується на вивченні ефективних методів автоматизованого розпізнавання дезінформації, використовуючи передові техніки нейронних мереж та обробки природної мови. Особливий акцент робиться на виявленні емоційного впливу, який часто супроводжує дезінформаційний контент. У роботі застосовано інноваційні підходи до аналізу текстів з метою виявлення прихованих технік маніпулювання емоційним станом аудиторії, які застосовуються поширювачами неправдивої інформації. Результатом дослідження є розробка високоточних систем розпізнавання та категоризації емоційного забарвлення контенту, пов'язаного з дезінформацією. Такі системи здатні ефективно виявляти та фільтрувати потенційно шкідливі матеріали, тим самим посилюючи стійкість інформаційного середовища. Запропоновані рішення можуть стати важливим внеском у боротьбу з поширенням дезінформації та сприяти підвищенню рівня кібербезпеки, забезпечуючи надійність та цілісність інформаційного простору. Отримані результати мають потенціал для вдосконалення засобів протидії дезінформації та зміцнення довіри до інформаційного середовища. Дослідження включає аналіз етичних аспектів застосування нейронних мереж для розпізнавання дезінформації та розробку відповідних стандартів, спрямованих на захист приватності користувачів. Комплексна робота пропонує надійну систему фільтрації та реагування на потенційні дезінформаційні загрози, відкриваючи нові перспективи для підвищення рівня кібербезпеки та забезпечення надійності інформаційного середовища.*

Ключові слова*: розпізнавання дезінформації; неправдиві новини; методи виявлення неправдивої інформації та фейкових новин в Інтернеті.*

Dmytro Rabchun
State University of information and commucation technogies, Kyiv
ORCID ID: 0000-0002-5555-0910

Vitalii Tyshchenko
State University of information and commucation technogies, Kyiv
ORCID ID: 0000-0003-3849-6243

Serhii Goloborodko
State University of information and commucation technogies, Kyiv
ORCID ID: 0009-0006-9892-3031

EFFECTIVE RECOGNITION OF MISINFORMATION WITH THE HELP OF NEURAL NETWORKS: FOCUS ON THE DETECTION OF EMOTIONAL IMPACT

***Abstract**. The research focuses on the study of effective methods for automated disinformation recognition using advanced neural network and natural language processing techniques. Particular emphasis is placed on identifying the emotional impact that often accompanies disinformation content. The study applies innovative approaches to text analysis to identify hidden techniques for manipulating the emotional state of the audience used by disseminators of false information. The result of the study is the development of high-precision systems for recognising and categorising the emotional colouring of disinformation-related content. Such systems are able to effectively detect and filter potentially harmful materials, thereby strengthening the resilience of the information environment. The proposed solutions can make an important contribution to the fight against the spread of disinformation and contribute to improving cybersecurity by ensuring the reliability and integrity of the information space. The findings have the potential to improve the means of countering disinformation and strengthening trust in the information environment. The study also includes an analysis of the ethical aspects of using neural networks to recognise disinformation and the development of appropriate standards aimed at protecting user privacy. This comprehensive work offers a robust system for filtering and responding to potential disinformation threats, opening up new perspectives for improving cybersecurity and ensuring the reliability of the information environment.*

***Keywords**: disinformation; fake news; methods of detecting disinformation and fake news on the Internet.*

1. Вступ.

Фейкові новини та дезінформація стали значною проблемою в сучасному суспільстві, а їхнє виявлення - важливим завданням. Одним із підходів до виявлення фейкових новин є використання нейронних мереж, зокрема, фокусуючись на емоційному впливі інформації. Аналіз емоцій відіграє ключову роль у визначенні поведінки користувача щодо певної теми, а фейкові новини мають навмисну мету збудити емоції читачів, щоб їм повірили.

Кілька досліджень показали, що фейкові новини викликають у людей страх, огиду і здивування. Тому виявлення емоційного впливу інформації може бути ефективним способом виявлення фейкових новин. Одним із типів нейронних мереж, які можуть бути використані для цього завдання, є згорткові нейронні мережі (CNN), які використовують шари згортання для відображення особливостей вхідних даних. Однак, ефективність цих методів залежить від якості та кількості навчальних даних.

У зв'язку з цим дослідники запропонували різні підходи, зокрема подвійні емоційні ознаки, нейронні мережі на основі графів та аналіз настроїв, щоб покращити виявлення фейкових новин.

2. Постановка проблеми.

Проблема полягає у виявленні фейкових новин та дезінформації за допомогою нейронних мереж з акцентом на виявленні емоційного впливу інформації. Однак ефективність цих методів залежить від якості та кількості навчальних даних. Дослідники запропонували різні підходи, зокрема використання згорткових нейронних мереж (CNN), функцій подвійних емоцій, нейронних мереж на основі графів та аналізу настроїв, щоб покращити виявлення фейкових новин. Однією з ознак дезінформації є її емоційний вплив. Дезінформація часто розробляється таким чином, щоб викликати у читачів емоції, такі як страх, гнів або ненависть. Ці емоції можуть зробити людей більш сприйнятливими до дезінформації.

3. Аналіз останніх досліджень і публікацій.

Останні наукові публікації концентруються на застосуванні нейронних мереж для виявлення фейкових новин та дезінформації, приділяючи особливу увагу аналізу емоційного впливу інформації. Для підвищення ефективності виявлення, дослідники запропонували різні методологічні підходи, серед яких - використання згорткових нейронних мереж (CNN), методів аналізу подвійних емоційних ознак, нейронних мереж sentiment-аналізу [1].

Аналіз емоційного забарвлення інформації відіграє ключову роль у визначенні поведінки користувачів щодо певної тематики, оскільки фейкові новини навмисно спрямовані на

збудження емоцій читачів з метою посилення їх довіри. Отже, виявлення емоційного впливу може бути ефективним підходом до ідентифікації фейкового контенту [2]. Водночас результативність таких методів значною мірою залежить від якості та обсягу навчальних даних, доступних дослідникам.

Наукові колективи запропонували різні інноваційні рішення для вдосконалення виявлення фейкових новин, зокрема - використання подвійних емоційних ознак, застосування нейронних мереж на основі графів [3].

Одне з репрезентативних досліджень порівнює ефективність чотирьох моделей машинного навчання, включаючи згорткові нейронні мережі (CNN) та інші типи нейронних архітектур, у задачі розпізнавання та протидії дезінформації [4].

Інше дослідження аналізує сучасні методи автоматизованого розпізнавання емоцій, зокрема, використання методу Віоли – Джонса для виявлення емоційних ознак на основі прямокутних характеристик [5]. Це свідчить про стійкий науковий інтерес до застосування нейронних мереж у сфері аналізу емоційного складника інформації.

4. Мета дослідження полягає в розгляді сучасних методів розпізнавання емоцій та їх застосування для виявлення дезінформації.

Для досягнення цієї мети необхідно вирішити наступні завдання:

Аналіз останніх досліджень та публікацій щодо використання нейронних мереж для розпізнавання емоцій в інформації.

Вивчення ефективності методів розпізнавання емоційного впливу в інформації з використанням нейронних мереж.

Аналіз впливу якості та обсягу тренувальних даних на ефективність розпізнавання емоцій за допомогою нейронних мереж.

Визначення перспектив використання нейронних мереж для виявлення емоційного впливу в інформації та застосування для боротьби з дезінформацією.

5. Виклад основного матеріалу.

Протягом останніх років соціальні медіа стали невід'ємною частиною життя багатьох людей і основним джерелом отримання новин. Зручний доступ, низька вартість та швидке поширення інформації спонукають користувачів використовувати соціальні мережі для отримання новин [6, с.49]. Однак це також призводить до поширення недостовірної інформації, включаючи фейкові новини, що може негативно впливати на суспільство.

Фейкові новини у соціальних мережах, такі як підроблені відео та блоги, створені анонімними особами, представляють собою серйозну загрозу громадськості. Виявлення фейкової інформації стало окремою галуззю досліджень, оскільки їх швидке поширення може спричинити негативні наслідки для суспільства. Розповсюдження фейкових новин може бути зумовлене необхідністю привернення уваги аудиторії через конкурентне середовище соціальних мереж [7]. Користувачі, які швидко споживають інформацію, стають більш вразливими перед впливом недостовірних даних. Відсутність раціональних основ та логіки у публікаціях дозволяє інтернет-ресурсам залишатися безкарними при поширенні неправдивої інформації. В епоху, позначену швидким розповсюдженням інформації через цифрові канали, проблема розрізнення справжнього контенту від дезінформації стає дедалі складнішою. Традиційні методи перевірки фактів і перевірки не встигають за обсягом і швидкістю потоку інформації.

Поширення дезінформації через соціальні медіа та Інтернет загрожує суспільному добробуту. Розпізнавання емоційного впливу важливо для виявлення дезінформації і нейронні мережі можуть допомогти аналізувати емоційний підтекст контенту. Маніпуляція емоціями грає ключову роль в успішності кампаній з дезінформації. Тому використання нейронних мереж допомогає обробляти великі обсяги даних та розпізнавати шаблони для виявлення емоційних складових у контенті.

Нейронні мережі, особливо моделі глибокого навчання, продемонстрували чудові можливості в різних задачах обробки природної мови. Ці моделі можна навчити розпізнавати шаблони в мові, які вказують на емоційний вплив. Використовуючи великі набори даних, які охоплюють як законний, так і дезінформаційний вміст, нейронні мережі можуть навчитися ідентифікувати тонкі лінгвістичні ознаки, пов’язані з емоційними маніпуляціями.

Не дивлячись на потенціал нейронних мереж у виявленні емоційних афектів, існує низка проблем, які ускладнюють їх ефективне застосування. Динаміка мови, культурні відтінки та зростаюча кількіть дезінформаційного контенту вимагають надійних моделей, що можуть адаптуватися до нових стратегій. Більше того, етичні аспекти управління емоційно наповненим контентом та можливі упередження в навчальних наборах даних потребують уважного розгляду. Для підвищення ефективності виявлення дезінформації на основі нейронних мереж необхідний міждисциплінарний підхід. Співпраця між лінгвістами, психологами та комп’ютерними фахівцями може забезпечити повне розуміння емоційних та психологічних механізмів. Шляхом об'єднання досвіду в цих галузях ми можемо вдосконалити моделі нейронних мереж для кращого виявлення найтонших відтінків емоційного впливу, що вказують на дезінформацію.

Як і в усіх технологічних здобутках, застосування нейронних мереж для розпізнавання дезінформації породжує етичні питання. Питання, пов'язані з конфіденційністю, свободою висловлювання думок та потенційними упередженнями алгоритмів, потребують уважного розгляду. Забезпечення балансу між необхідністю боротьби з дезінформацією та захистом індивідуальних прав потребує розробки етичних принципів і нормативної бази. Інтеграція мультимодальних підходів, які комбінують текстовий та візуальний аналіз, а також дослідження механізмів виявлення у реальному часі, мають захоплюючі перспективи. Крім того, вирішення проблеми інтерпретації моделей нейронних мереж може підвищити їх достовірність і сприяти співпраці між аналітиками та автоматизованими системами.

У широкому розумінні термін "фейк" визначає будь-яку подійну або предметну підробку, яку особа намагається представити як автентичну. Наприклад, фотографії НЛО часто класифікуються як фейки, проте перевірка їхньої автентичності не завжди можлива. Завдяки сучасним технологіям, поширення фейків стає досить поширеним явищем, що створює складність у їхній перевірці. Фейками можуть бути акаунти у соціальних мережах, інтернет-сайти, підробки відомих брендів, фармацевтичні препарати, товари у роздрібних магазинах та інше. Проте, серед всіх цих категорій, особливо важливими є фейкові новини, що поширюються через ЗМІ.

Виявлення фальшивих новин стає надзвичайно важливим завданням, що залучає все більше уваги, оскільки фейки негативно впливають на суспільство. Однак ефективність виявлення фейків за їхнім змістом часто залишає бажати кращого. Для поліпшення цього процесу необхідні спеціальні методи та інструменти, які базуються на вивченні взаємозв'язку між характеристиками фейкової інформації та поведінкою користувачів у соціальних мережах [8].

Фотофейки є одним із найбільш поширених типів дезінформації, і їх легко спростувати. Існує багато способів виявлення фотофейків. Наприклад, можна клацнути правою кнопкою миші на незнайомому зображенні в браузері Google Chrome і вибрати опцію «Шукати це зображення в Google». Для інших браузерів, де за замовчуванням ця функція відсутня, можна встановити спеціальні плагіни, такі як «Хто вкрав мої фотографії». Цей плагін значно підвищує ефективність пошуку фотофейків за допомогою Google, Tineye та інших пошукових систем одночасно.

У випадку відеофейку виявлення є складнішим, оскільки прямого способу пошуку відео немає. Якщо у вас є підозри щодо відео, спробуйте деякі наступні методи. По-перше, перейдіть на сам YouTube, якщо ви дивитеся вбудоване вікно на іншому сайті, щоб отримати більше інформації про відео. Якщо немає очевидних ознак відеофейка, зверніть увагу на деякі деталі. Нова дата в назві відео та багаторазове завантаження на YouTube протягом короткого періоду часу можуть свідчити про його неправдивість. Розгляньте відео з найбільшою кількістю

переглядів і перегляньте коментарі - це може допомогти визначити, чи мали справу з оригінальним відео. Також обережно вивчайте деталі на відео, такі як назви об'єктів, автомобільні номери, вуличні таблички, оскільки це може вказати на справжність чи фейковість відео.

Фальшивий журналістський матеріал часто використовує такі прийоми, як посилання на авторитетні ЗМІ або перекручення повідомлень та коментарів для зміцнення авторитетності та правдоподібності інформації. Цей метод сприяє створенню враження, що неправдивий матеріал має підтримку від відомих джерел. Проте, серед таких повідомлень може бути багато вигадок маргінальних сайтів, що потребує перевірки. Фальшиві журналістські матеріали іноді спеціально створюються навколо вигаданих новин, інсценізуючи їх.

Замість прямої боротьби з фейками, рекомендується перевіряти їх та критикувати. Важливо уникати поширення фейків, оскільки це може призвести до "інформаційного шуму", який відволікає від важливої інформації. Боротьба із фейками передбачає виявлення та ігнорування неправдивих повідомлень, сприяючи їхньому природному відторгненню.

Існує кілька способів відрізнити фейк в новинах*:*

1. Аналіз джерела інформації. Якщо повідомлення базується виключно на повідомленнях із соціальних мереж, без підтвердження з інших джерел, слід ставитися до нього з обережністю. Наприклад, інформація про масове піднімання літаків, що походить лише з одного твіту, потребує ретельної верифікації.
2. Оцінка характеру заголовків та формулювань. Новини чи повідомлення з сенсаційними заголовками або риторичними запитаннями можуть бути спробами маніпулювати увагою читача та створювати ілюзію важливості чи ексклюзивності інформації.
3. Врахування емоційного забарвлення. Нейтральні, фактологічні новини мають більше шансів бути достовірними. Натомість, інформація з явними негативними емоційними елементами, використанням епітетів та ярликів, може свідчити про наявність пропаганди.
4. Перевірка автентичності джерел. Фейкові матеріали можуть розповсюджуватись під виглядом справжніх акаунтів. Тому варто звертати увагу на нещодавно створені акаунти з мінімальною активністю та запозиченими фотографіями, особливо у спробах тематичних маніпуляцій.

За допомогою зазначених підходів можна провести більш обґрунтований аналіз інформації та оцінити її достовірність.

Існуючі алгоритми виявлення фейків у традиційних мас-медіа часто виявляються неефективними у роботі з фейками у соціальних мережах. Традиційні ЗМІ зазвичай докладно перевіряють інформацію, оскільки вони несуть відповідальність перед вповноваженими організаціями. Отже, поява неправдивих або сумнівних новин у традиційних ЗМІ може миттєво пошкодити їхню репутацію та призвести до припинення їхньої діяльності. У той же час, соціальні мережі, де кожен може публікувати контент, все частіше стають платформою для поширення фейків, які сприймаються як норма.

Деякі медійні видання можуть поширювати неперевірену або неправдиву інформацію з метою привернення більшої уваги до своїх ресурсів або впливу на громадську думку. Різноманітні особи можуть розповсюджувати інформацію в Інтернеті, що може призвести до значного зростання кількості фейкових повідомлень. Це явище не обов'язково має виникати внаслідок недорозумінь, але часто має виражену мету впливу на громадську думку або відволікання уваги від ключових подій. Підкреслюється, що навіть в разі випадкових інформаційних помилок спостерігається зростання тенденції до умисного поширення неправдивої інформації.

Фейки розрізняються за метою їх створення та поширення. Основні типи фейків включають:

- Створення паніки серед населення;
- Розпалювання міжнаціональної, расової або релігійної ворожнечі;
- Поширення хибних думок для заплутування та відволікання від правди;

- Маніпулювання свідомістю;
- Реклама певних осіб або об'єктів;
- Генерування прибутку для медіа ("жовта преса");
- Очорнення репутації через підроблені фотографії;
- Розважальний контент.

Неправдива інформація шириться в цифровому просторі шляхом коментарів, переважно з фальшивих акаунтів, а також за допомогою хештегів, особливо під час популярних годин, коли користувачі активно шукають інформацію за певними тегами. В цей період спамери ініціюють інформаційні атаки, використовуючи зазначені хештеги. Автори фейкових новин та іншої неправдивої інформації часто намагаються імітувати цитування авторитетних джерел, поширюючи скріншоти з видимими цитатами відомих осіб. Ця інформація може здаватися дуже переконливою, але не є достовірною.

У таких випадках завжди важливо перевіряти наявність цієї інформації на офіційному веб-сайті або офіційній сторінці особи, на яку посилаються. Також варто відзначити тенденцію до поширення неправдивої інформації офіційними ЗМІ, які стали ключовим інструментом для впливу на громадську свідомість.

Річард Ніксон, колишній президент США, виступаючи перед Радою національної безпеки США, висловив думку про важливість інвестування в інформацію та пропаганду. Він зазначив, що кожен долар, витрачений на ці цілі, має більший вплив, ніж коштів, витрачених на розбудову військово-промислового комплексу, оскільки інформація має непередбачуваний та миттєвий вплив на суспільство [10].

Дезінформаційний контент має потенціал відволікати увагу реципієнтів від критично важливих інформаційних потоків. Аналіз функціональності фальшивих новин дозволяє виділити три основні аспекти їх впливу:

- Підвищення рейтингу популярності або монетизація діяльності суб'єктів, що генерують такий контент;
- Перенасичення інформаційного простору для аудиторії, яка критично сприймає подібні повідомлення, але все ж піддається їх впливу через постійну експозицію;
- Маніпулятивний вплив на когнітивні процеси довірливої частини аудиторії, що некритично сприймає даний тип інформації.

Зазначені характеристики є фундаментальним підґрунтям для розробки ефективних механізмів протидії дезінформації. Застосуванню методів детекції фейкових новин має передувати комплексний попередній аналіз інформаційного контенту.

У контексті соціальних медіа, фейковий новинний контент часто характеризується високою концентрацією емоційно забарвленої лексики та потенційно образливих висловлювань. Наявність такого лексичного профілю є суттєвим індикатором потенційної недостовірності інформації, що обумовлює необхідність більш ретельної верифікації.

Лінгвістичний аналіз тексту, зокрема оцінка емоційної тональності та виявлення специфічних лексичних патернів, може слугувати первинним фільтром для ідентифікації потенційно недостовірного контенту. Це дозволяє оптимізувати процес подальшої верифікації, зосереджуючи ресурси на найбільш підозрілих інформаційних одиницях [11].

Методологія детекції дезінформації характеризується диференціацією підходів залежно від типології аналізованого контенту. Основними категоріями контенту, що підлягають верифікації, є текстовий, відео- та аудіоматеріали, кожен з яких вимагає специфічних алгоритмів аналізу.

У контексті верифікації текстового контенту одним із фундаментальних інструментів ідентифікації недостовірної інформації є метод фактчекінгу. Цей підхід базується на систематичній верифікації фактологічної основи повідомлення шляхом звернення до первинних джерел інформації.

Якщо необхідно виявити фейк у відеоконтенті, першим кроком є пошук першоджерела. Однак у цьому випадку ідентифікація використаних джерел може бути складнішою. Штучний

інтелект здатен допомогти у виявленні недостовірної інформації у відео, і дедалі частіше застосовується для таких цілей. Під час аналізу достовірності інформації у відеоконтенті, штучний інтелект звертає увагу на такі лінгвістичні прийоми, як надмірно емоційні, гіперболізовані чи образливі висловлювання [12, с.64].

Штучний інтелект є одним із найефективніших засобів для виявлення недостовірної інформації у медіаконтенті. Відповідне програмне забезпечення, що ґрунтується на технологіях штучного інтелекту, неодноразово підтверджувало свою дієвість у цій сфері [13, с.66].

Архітектура такого програмного забезпечення базується на моделі довгої короткочасної пам'яті, що є найбільш придатною для обробки природної мови. Завдяки цій моделі, штучний інтелект встановлює зв'язки між послідовними словами, словосполученнями та реченнями, а потім аналізує достовірність наведеної інформації в контексті. Як правило, в якості інформативних ознак використовують заголовки, ключові слова чи початкові слова тексту новин, на основі яких формується вибірка даних. За допомогою цієї вибірки, новина класифікується як достовірна чи недостовірна.

Виявлення недостовірних новин стикається з трьома значущими проблемами. По-перше, платформи недостовірних новин часто імітують справжні новинні ресурси як за дизайном, так і за контентом. По-друге, недостовірні новини не завжди є абсолютною істиною, і в них можуть міститися істинні факти, перемішані з хибними твердженнями. По-третє, сучасні технології дозволяють фабрикувати зображення та відео, ускладнюючи миттєву перевірку правдивість новин. Відтак, створення недостовірних суб'єктів для поширення недостовірних новин стає простим завдяки безкоштовним інструментам в Інтернеті, а також можливості підміни облич за допомогою наявних фотографій.

Статті фейкових новин, як правило, мають романтичний або драматичний зміст і часто використовують більше порівнянь та емоційно забарвлених слів для гри на почуттях, ніж справжні новини. Отже, для виявлення фейкових новин необхідна ефективна обробка природної мови (NLP). Однак NLP є складним та трудомістким процесом через різні значення слів у різних контекстах, а також потребує окремих версій для кожної мови та діалекту. Недостатня кількість доступних даних, а також вища кількість справжніх новин порівняно з фейковими, обмежують ефективність навчання моделей класифікації.

Раніше класифікаційні та регресійні моделі, які вивчали або зміст статті, або канали поширення чуток, допомогли виявити фейкові новини. Однак необхідність більш ефективних методів спонукала до впровадження моделей глибоких нейронних мереж, зокрема двонаправленої довготривалої короткочасної пам’яті (LSTM).

Ця модель використовує різноманітні вхідні дані для ефективного виявлення фейкових новин. Вміст статті представлений у вигляді вкладених слів, що дозволяє краще розуміти смисл інформації. Метадані включають описові параметри для статей, твітів та ретвітів, такі як ідентифікатори, контент і кількість підписників. Описові параметри в метаданих є числовими атрибутами, нормалізованими та дискретизованими.

Часові ряди створюються на основі шляхів поширення чуток, що дозволяє враховувати динаміку розповсюдження інформації. Усі ці вхідні дані інтегруються та вводяться до двонаправленої мережі LSTM, яка класифікує новини як справжні або фейкові.

Для оцінки ефективності цієї моделі були розроблені та випробувані інші методи, такі як CNN, C-LSTM (з використанням метаданих і шляхів розповсюдження чуток), CNN+LSTM (що поєднує зміст статті, метадані та шляхи розповсюдження чуток), Multinomial NB, SVM, а також моделі з попередніх досліджень. Запропонована модель покращує точність виявлення фейкових новин завдяки вдосконаленій архітектурі та інтеграції різних типів вхідних даних.

За останні роки системи виявлення фейкових новин значно розвинулися в боротьбі з поширенням дезінформації в Інтернеті. Застосування методів машинного навчання, особливо різноманітних класифікаторів, які продемонстрували високу ефективність у виконанні цього завдання, є важливою частиною таких систем. Метод опорних векторів (SVM), метод k-

найближчих сусідів (KNN), дерева рішень і випадковий ліс (RF) є одними з можливих алгоритмічних підходів.

Дослідження, проведене на наборі даних від Kaggle Fake News Challenge, показало, що найвища точність досягається за допомогою комбінації алгоритмів Natural Language Processing (NLP) та класифікатора Random Forest.

Це свідчить про те, що у даному випадку класичні алгоритми машинного навчання виявилися ефективнішими, ніж моделі глибокого навчання. Є кілька можливих причин для цього, на (рис.1.) розглянуто складові машинного навчання.

Рис. 1. Складові машинного навчання

Останнє десятиліття характеризується інтенсифікацією наукових досліджень, спрямованих на вивчення феномену довіри до систем штучного інтелекту (ШІ). Значна когорта видатних представників наукової та технологічної спільноти, зокрема Стівен Хокінг, Ілон Маск та Білл Гейтс, артикулюють глибоку стурбованість щодо потенційних ризиків, асоційованих з прогресом високорозвинених систем ШІ та їх імплікацій для соціуму. Ця проблематика набуває особливої актуальності в контексті стрімкого розвитку та впровадження технологій ШІ в різноманітні сфери людської діяльності.

У дослідженні [14] науковці пропонують концептуальну модель для аналізу та вирішення проблематики довіри до систем штучного інтелекту (ШІ). Ключовими компонентами довіри визначено транспарентність, етичні аспекти, технічну надійність та захист конфіденційності даних. Запропонована модель інтегрує такі виміри, як індивідуальні характеристики, технологічні параметри та контекстуальні фактори. Це дозволяє здійснювати комплексний аналіз інтеракцій між людиною та ШІ, досліджувати етичні аспекти процесів прийняття рішень та оцінювати вплив різних рівнів розвитку ШІ на формування довіри до цих систем.

Додатково, в 2023 році були опубліковані результати глобального дослідження "Довіра до штучного інтелекту", яке детально аналізує рівень довіри та ставлення громадськості до ШІ у 17 країнах. Звіт висвітлює очікування спільноти, регулювання та управління ШІ, а також надає важливу інформацію для формування політики та стратегій використання ШІ в різних галузях, включаючи бізнес, уряд та неурядові організації.

З даних глобальне опитування у 28 країнах стосовно загальних думок та очікувань щодо штучного інтелекту. Результати свідчать, що більшість людей у світі чули про ШІ, але лише невелика їх частка розуміє його можливості. Багато людей ставляться до ШІ позитивно, хоча є й ті, хто висловлює стурбованість його можливими негативними наслідками. Довіра до ШІ є відносно низькою, а ключовими чинниками, що на неї впливають, є прозорість, зрозумілість та підзвітність. Більшість людей вважає, що ШІ потребує належного регулювання для забезпечення відповідального використання [15].

В подальших дослідженнях може бути важливо враховувати особливості наборів даних та оптимізувати параметри моделей для покращення результатів. Також, розвиток методів глибокого навчання може призвести до подальшого підвищення ефективності систем виявлення фейкових новин.

Дослідження ефективності реальних методів класифікації у завданнях машинного навчання акцентується на використанні різноманітних класифікаторів, таких як логістична регресія, модель LDA, модель QDA, KNN-модель, дерево рішень та випадковий ліс, з метою виявлення фейкових новин. В ході дослідження використовується комбінація даних з набору "Liar liar pants on fire" та участь у змаганні з фейкових новин Kaggle Fake News. Розроблена методологія класифікації представлена у вигляді послідовності етапів, як зображено на рисунку 2.

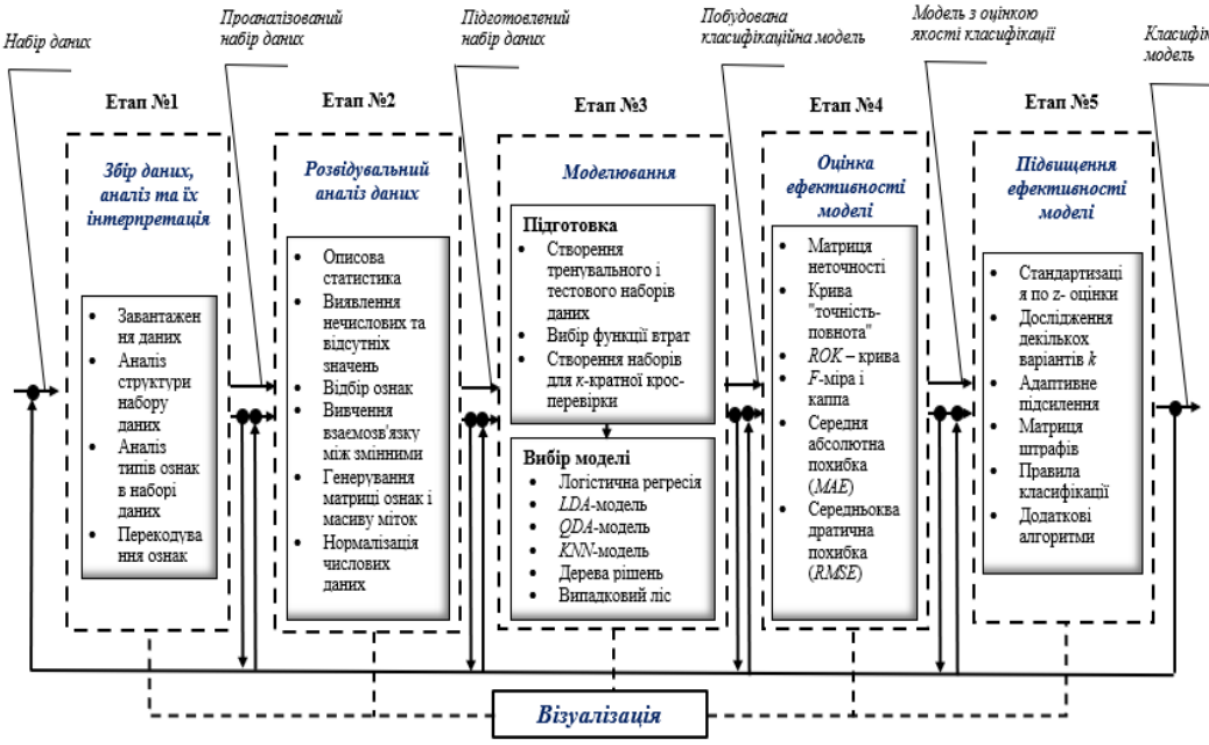

Рис. 2. Послідовність етапів вирішення завдань класифікації

Основною передумовою для поширення недостовірних новин є існування Інтернету та соціальних мереж як основних платформ, на яких більшість людей отримує свіжі новини та інформацію. За даними DataReportal на 2024 рік, кількість користувачів Інтернету становить 5,32 мільярда, а соцмереж - 4,65 мільярда. Для успішного поширення фейків необхідні інструменти та послуги для маніпулювання та поширення неправдивих повідомлень, які доступні в різних онлайн-спільнотах по всьому світу [16].

В інформаційному просторі існує різноманітність інструментів і послуг, спрямованих на поширення фейкових новин. Деякі з них є відносно простими і доступними, наприклад, платні послуги для отримання лайків та підписників у соціальних мережах, тоді як інші є більш складними та нестандартними. Деякі сервіси пропонують проведення онлайн-опитувань, а інші змушують власників веб-сайтів видаляти статті або історії, що можуть розголошувати правдиву інформацію. Ці інструменти та послуги доступні як у хакерських спільнотах, так і поза ними, що підвищує їх доступність та розповсюдження.

Ключовим елементом поширення фейкових новин є мотивація їхніх поширювачів. Це може бути спонукання до фінансової вигоди через рекламу або досягнення різноманітних цілей, включаючи політичні та кримінальні мотиви. Незалежно від мотивації, успішність дезінформації визначається її впливом на реальний світ. З урахуванням величезного розмаху дезінформації в глобальному інтернет-просторі та постійного зростання її масштабів, виникає

проблема необхідності запобігання та протидії цьому явищу як інструменту маніпулювання громадською думкою і суспільною свідомістю.

Запропоновані підходи включають комбінацію традиційних принципів критичного мислення з використанням сучасних технологій, зокрема алгоритмів нейронних мереж.

1. Перевірка джерела новин:
   - Використання нейромережевих алгоритмів для автоматичної перевірки легітимності та надійності джерел.
   - Аналіз дотримання стандартів точності, збалансованості та об'єктивності за допомогою нейромережевих моделей.
2. Аналіз змісту та заголовків:
   - Використання алгоритмів для виявлення упередженості презентації, сенсаційних заголовків та перекручених даних.
   - Автоматична критична оцінка контенту з використанням нейромережевих алгоритмів для ідентифікації непідтверджених джерел та статистичних даних.
3. Перевірка інформації про автора:
   - Використання алгоритмів для оцінки надійності автора на основі аналізу його досвіду, репутації та попередньої діяльності.
   - Автоматизоване визначення авторитетності статей за допомогою нейронних мереж та професійної біографії автора.
4. Перевірка посилань на новини:
   - Застосування нейромережевих алгоритмів для аналізу інформації про осіб, згаданих у новинах, з оцінкою їхньої кваліфікації та об'єктивності.
   - Автоматичне виявлення відсутності належних посилань або сумнівності анонімних джерел.
5. Перевірка актуальності новин:
   - Використання алгоритмів для підтвердження своєчасності новин шляхом аналізу дати та часу публікації та порівняння з іншими джерелами.

Застосування нейромережевих технологій може суттєво підвищити ефективність виявлення фейкових новин та сприяти зростанню довіри користувачів до інформації, що поширюється в соціальних мережах. Водночас, основні навички критичного мислення та обачності залишаються ключовими у протидії дезінформації навіть за використання передових технологічних рішень.

6. Висновки та перспективи подальших досліджень.

Розвиток технологій у сфері штучного інтелекту, зокрема застосування глибокого навчання в нейронних мережах, стає ключовим фактором у боротьбі з поширенням дезінформації. Емоційний вплив виявляється надзвичайно важливим у формуванні поглядів та переконань людей, і відомо, що цей фактор активно використовується в метах дезінформації.

Доведено, що нейронні мережі можуть успішно розпізнавати емоційні сигнали в текстах, що допомагає в ідентифікації потенційно обманливих матеріалів. Використання глибокого навчання дозволяє підвищити здатність мереж розрізняти тонкості емоційного впливу, що забезпечує точніше визначення можливих випадків дезінформації.

Застосування контекстуального аналізу, яке враховує ситуаційний контекст, може покращити ефективність розпізнавання дезінформації. Розгляд відмінностей культур та залучення міжнародної спільноти у подальші дослідження дозволить розширити глобальний підхід до цієї проблеми.

Розширення обсягу досліджень на інші типи медіа та соціальні мережі допоможе створити більш комплексні моделі для розпізнавання дезінформації. Розробка систем для реального часу, які можуть виявляти дезінформацію миттєво, дозволить ефективно реагувати на нові загрози. Оптимізація взаємодії з користувачами та розробка зручних інтерфейсів для перевірки достовірності інформації будуть ключовими елементами протидії дезінформації.

Список використаних джерел

1. Сучасні підходи до виявлення та протидії дезінформації в інформаційн - львівська політехніка URL: https://science.lpnu.ua/sites/default/files/journal-paper/2023/nov/32213/10.pdf
2. Метод нейромережевого розпізнавання фальсифікованих зображень URL: http://bionics.nure.ua/article/download/228459/227559/519971
3. Розпізнавання емоцій людини за допомогою згорткової нейронної мереж URL:https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/30837/69737.pdf?isallowed=y&sequence=2
4. Deep learning: основи напряму, інструменти та технології URL: https://foxminded.ua/deep-learning/
5. Нейромережева модель розпізнавання емоцій по зображенню обличчя URL: http://www.tech.vernadskyjournals.in.ua/journals/2019/2_2019/part_1/35.pdf
6. Бахтіна Г.П. Інформатизація суспільства та проблема «кліпового мислення». Київ, 2011. URL: https://kpi.ua/1102-7
7. Денніс А., Моравец П. та Кім А. (2023). Пошук і перевірка: дезінформація та оцінки джерел у результати пошуку в Інтернеті. Системи підтримки прийняття рішень, 171. URL: https://doi.org/10.1016/j.dss.2023.113976
8. Заброда В.Є., Льовкін В.М. Програмне забезпечення розпізнавання неправдивих новин. Free and open source software: Харківська ювілейна міжнародна науково-практична конференція (Харків, 20–22 листопада 2018). - Харків, 2018. С.66.
9. Іванова І., Лисицька О. Постмодернізм як маніпулятивна технологія в сучасній українській рекламі: характеристика художньої домінанти (Стаття). Міжнародний філологічний часопис. Серія: Соціальні комунікації. том. 11, № 1; 2020. С. 108-113. URL: http://dx.doi.org/10.31548/philolog 2020.01.108
10. Кардуні, А. Взаємодія людини та дезінформації: розуміння необхідного міждисциплінарного підходу для комп'ютерної боротьби з неправдивою інформацією. Цифрова бібліотека ACM. URL: https://doi.org/10.1145/1122445.1122456
11. Лісневська А. Дезінформація в новинному відеоконтенті: маркери та методи розпізнавання. Вісник Львівського університету. 2019. № 45. С.60–66.
12. Почепцов Г. (Дез)інформація. Бібліотека ГО "Детектор медіа". За спільною редакцією Н. Лігачової та Г. Петренка / Редакцією М. Олійник. - Київ, 2019. С.248.
13. Растогі, С., і Бансал, Д. (2023). Огляд виявлення фейкових новин 3Т: типологія, час виявлення, таксономії. Міжнародний журнал інформаційної безпеки. URL: https://doi.org/10.1007/s10207-022-00625-3
14. Ройтер К., Гартвіг К., Кірхнер Дж. та Шлегель Н. Сприйняття фейкових новин у Німеччині: репрезентативне дослідження ставлення людей і підходів до протидії дезінформації. Wirtschaftsinformatik. 2021, С.54- 61.
15. Сибиряков С. Соціальні медіа як середовище архетипового впливу на масову свідомість. Політичне управління: теорія і практика. 2013. № 1. С. 202– 210.
16. Татарчук Д.О. Інструменти фактчекінгу при виявленні фейкової інформації в соціальних медіа. International scientific and practical conference (Влоцлавек, Польща, 27–28 листопада 2020). Влоцлавек, - 2020. С. 84–86.
17. Lukyanenko R., Maass W., Storey V. C. Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electronic markets*. 2022. URL: https://doi.org/10.1007/s12525-022-00605-4
18. Бейбкок М., Бесков & Д. Карлі К.. (2018). Різні обличчя фальші: поширення та Зменшення неправдивої інформації в обговоренні в Твіттері Чорної Пантери. Якість даних та інформації. URL: https://doi.org/10.1145/3339468
19. DataReportal. URL: https://datareportal.com/reports/digital-2024-deep-dive-5-billion-social-media-users?rq=users

References

1. Modern approaches to detecting and counteracting disinformation in the information sector - Lviv Polytechnic URL: https://science.lpnu.ua/sites/default/files/journal-paper/2023/nov/32213/10.pdf
2. Method of neural network recognition of falsified images URL: http://bionics.nure.ua/article/download/228459/227559/519971
3. Recognizing human emotions using convolutional neural networks URL: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/30837/69737.pdf?isallowed=y&sequence=2
4. Deep learning: basics, tools and technologies URL: https://foxminded.ua/deep-learning/
5. Neural network model for recognizing emotions from a face image URL: http://www.tech.vernadskyjournals.in.ua/journals/2019/2_2019/part_1/35.pdf
6. Bakhtina G.P. Informatization of society and the problem of "clip thinking". Kyiv, 2011. URL: https://kpi.ua/1102-7.
7. Dennis A., Moravec P. and Kim A. (2023). Search and verification: disinformation and source evaluations in Internet search results. Decision Support Systems, 171. URL: https://doi.org/10.1016/j.dss.2023.113976
8. Zabroda V.E., Levkin V.M. Fake news recognition software. Free and open source software: Kharkiv Jubilee International Scientific and Practical Conference (Kharkiv, November 20-22, 2018). - Kharkiv, 2018. C.66.
9. Ivanova I., Lysytska O. Postmodernism as a Manipulative Technology in Modern Ukrainian Advertising: Characterization of the Artistic Dominant (Article). International philological journal. Series: Social Communications. vol. 11, № 1; 2020. C. 108-113. URL: http://dx.doi.org/10.31548/philolog 2020.01.108
10. Carduni, A. Human interaction and disinformation: understanding the necessary interdisciplinary approach for computer-based countering false information. ACM Digital Library. URL: https://doi.org/10.1145/1122445.1122456
11. Disinformation in news video content: markers and methods of recognition. Bulletin of Lviv University. 2019. № 45. C.60-66.
12. Pocheptsov G. (Dis)information. Library of the NGO "Detector Media". Edited by N. Ligacheva and H. Petrenko / Edited by M. Oliynyk. - Kyiv, 2019. C.248.
13. Rastogi, S., and Bansal, D. (2023). A review of 3T fake news detection: typology, detection time, taxonomies. International Journal of Information Security. URL: https://doi.org/10.1007/s10207-022-00625-3
14. Reuter, K., Hartwig, K., Kirchner, J. and Schlegel, N. Perception of fake news in Germany: a representative study of people's attitudes and approaches to countering disinformation. Wirtschaftsinformatik. 2021, C.54- 61.
15. Social media as a medium of archetypal influence on the mass consciousness. Political management: theory and practice. 2013. № 1. C. 202- 210.
16. Tatarchuk D.O. Fact-checking tools for detecting fake information in social media. International scientific and practical conference (Włocławek, Poland, November 27-28, 2020). Włocławek, - 2020. C. 84-86.
17. Lukyanenko R., Maass W., Storey V. C. Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electronic markets*. 2022. URL: https://doi.org/10.1007/s12525-022-00605-4 .
18. Babcock M., Beskow & D. Carley C. (2018). Different Faces of Falsehood: Spreading and Mitigating False Information in the Black Panther Twitter Discussion. Data and Information Quality. URL: https://doi.org/10.1145/3339468
19. DataReportal. URL: https://datareportal.com/reports/digital-2024-deep-dive-5-billion-social-media-users?rq=users