МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ПОЛІСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет інформаційних технологій,
обліку та фінансів
Кафедра комп'ютерних технологій
і моделювання систем

Кваліфікаційна робота
на правах рукопису

Клименко Євген Сергійович

УДК 004.8:004.056

КВАЛІФІКАЦІЙНА РОБОТА

Технологія виявлення інформаційного впливу на акторів соціальних мереж

125 - Кібербезпека та захист інформації

Подається на здобуття освітнього ступеня магістр

кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ Клименко Є. С.

Керівник роботи
Тимонін Юрій Олександрович
К.т.н., доцент кафедри
комп'ютерних технологій і моделювання систем

Житомир – 2024

Висновок кафедри_____
за результатами попереднього захисту: _____

Протокол засідання кафедри _____
№ _____від «_____» _____20____р.

Завідувач кафедри _____

_____ _____ _____
(науковий ступінь, вчене звання) (підпис) (прізвище, ім'я, по батькові)
«_____» _____20____р.

Результати захисту кваліфікаційної роботи

Здобувач вищої освіти _____захистив (ла)
(прізвище ,ім'я, по батькові)
кваліфікаційну роботу з оцінкою:

сума балів за 100-бальною шкалою _____
за шкалою ECTS _____
за національною шкалою _____

Секретар ЕК

_____ _____ _____
(науковий ступінь, вчене звання) (підпис) (прізвище, ім'я, по батькові)

АНОТАЦІЯ

Клименко Є. С. Технологія виявлення інформаційного впливу на акторів соціальних мереж– Кваліфікаційна робота на правах рукопису.

Кваліфікаційна робота на здобуття освітнього ступеня магістр за спеціальністю 125 – кібербезпека та захист інформації. – Поліський національний університет,

Житомир, 2024.

Кваліфікаційна робота присвячена розробці та вдосконаленню методик виявлення інформаційного впливу на суб'єктів соціальних мереж за допомогою штучного інтелекту. Вивчає сучасні методи аналізу інформаційного впливу, зокрема з використанням інструментів машинного навчання та методіів обробки великих даних для аналізу контенту, ідентифікації вразливих груп користувачів та прогнозування можливих загроз.

У першому розділі роботи розглядаються теоретичні основи інформаційного впливу в соціальних мережах, обговорюється природа та основні механізми цього явища. Другий розділ охоплює огляд найсучасніших технологій і методів для виявлення явищ, що впливають на інформацію, зокрема щодо застосування штучного інтелекту. У третьому розділі пропонуються пропозиції щодо вдосконалення методів виявлення інформаційного впливу щодо ефективності методу штучного інтелекту та практичних потреб користувачів.

Основною частиною дослідження є побудова концептуальної моделі аналізу соціальних мереж, яка дозволяє ідентифікувати та розробляти стратегії проти інформаційних загроз. Цей підхід ґрунтуватиметься на гібридних формах аналізу даних, таких як аналіз тексту, кластеризація користувачів і прогнозування поведінки.

Отримані результати роботи можуть бути використані для покращення стану інформаційної безпеки та запобігання деструктивним впливам у соціальних мережах.

Загальна характеристика: кваліфікаційної роботи, 31 с., 10 дод., 11 джерел.

SUMMARY

Klymenko E. S. Technology for detecting information influence on social network actors – Qualification work on the rights of the manuscript.

Qualification work for the degree of master's degree in specialty 125 - cybersecurity and information protection - Polissya National University, Zhytomyr, 2023.

The qualifying work is devoted to the development and improvement of techniques for identifying informational influence upon actors in social networks through artificial intelligence. It studies the modern methods for the analysis of information influence, particularly with the usage of tools of machine learning and big data processing methods for content analysis, vulnerable groups' user identification, and predicting possible threats.

The first section of the paper deals with the theoretical foundations concerning information influence within social networks, with discussion of the nature and key mechanisms of the phenomenon. The second section covers an overview of state-of-the-art technology and techniques for the detection of information-influencing phenomena, particularly concerning artificial intelligence application. The third section proposes suggestions to improve the information-influence detection methods relative to efficiency in artificial intelligence method effectiveness and practical user needs.

The main part of the research is to construct a conceptual model of social network analysis that enables the identification and designing of strategies against information threats. The approach would be based on hybrid forms of data analysis like text analysis, clustering user, and prediction of behavior.

Results obtained from the work can be applied to improve the state of information security and prevent destructive influences in social networks.

General characteristics: qualification work, 31 p., 10 appendix, 11 sources.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AI (ШІ) – штучний інтелект

ML (МН) – машинне навчання

NLP – обробка природної мови

СМ (SM) – соціальні мережі

ІВ (II) – інформаційний вплив

DA – аналіз даних

IDS – системи виявлення вторгнень

OSINT – розвідка на основі відкритих джерел

RF – випадковий ліс

SVM – метод опорних векторів

LSTM – довга короткострокова пам'ять

BiLSTM – бінаправлена LSTM

API – інтерфейс прикладного програмування

IoT – Інтернет речей

КТіМС – комп’ютерні технології і моделювання систем

ВСТУП

Актуальність роботи Сучасна війна в Україні показала, що найважливішу роль у сучасних конфліктах відіграє інформаційна складова. Соціальні мережі (СМ) стали не лише засобом комунікації, а й чудовим інструментом формування громадської думки, налаштування настрою та маніпулювання інформаційним полем. У цих умовах виявлення та нейтралізація інформаційного впливу в СМ постає найбільш актуальною проблемою. Атаки інформаційного впливу, спрямовані на дестабілізацію суспільства, деморалізацію населення та дискредитацію державних інститутів, широко використовуються противником у стратегічних і тактичних цілях. Такі впливи часто здійснюються за допомогою дезінформації, фейків, бот-мереж, соціальної інженерії та інших методів маніпулювання. Однак через дуже високу швидкість поширення інформації в соціальних мережах ефективна протидія цим загрозам вимагає застосування передових технологій, зокрема штучного інтелекту (ШІ) та методів машинного навчання (МН). Інтеграція технологій ШІ дозволяє виявляти приховані шаблони інформаційної атаки, аналізувати величезні потоки даних у реальному часі, виявляти аномалії в поведінці користувача та прогнозувати можливі загрози. Це надає державним установам, волонтерським організаціям і громадянському суспільству ефективні інструменти проти інформаційних атак, одночасно поділивши можливий негативний ефект. У зв’язку з війною та посиленням залежності від цифрових технологій розробка та впровадження ефективних технологій виявлення впливу на інформацію є не просто науковою справою, а натомість критично необхідним елементом забезпечення національної безпеки України. Це підвищило важливість дослідження для створення ефективних методів для аналізу впливу інформації в соціальних мережах.

Мета роботи – розробка та впровадження технологій для виявлення та аналізу інформаційного впливу на акторів соціальних мереж.

Завдання:

– Проаналізувати існуючі технології виявлення інформаційного впливу в соціальних мережах.

– Розробити метод для систематизації та автоматизації збору даних з соціальних мереж.

– Оцінити ефективність методів виявлення інформаційного впливу у контексті протидії дезінформації та забезпечення кібербезпеки.

– Сформулювати рекомендації щодо використання технологій виявлення інформаційного впливу для підвищення рівня безпеки користувачів у соціальних мережах.

Об’єкт дослідження - процес виявлення інформаційного впливу на акторів соціальних мереж.

Предмет дослідження - технологія виявлення інформаційного впливу на акторів соціальних мереж.

Методи дослідження: аналіз наукової літератури та звітів, методи аналізу великих даних (Big Data Analytics), методи машинного навчання (ML), обробка природної мови (NLP), математичне моделювання, експериментальний метод.

Наукова новизна: полягає у комплексному підході до вирішення проблеми виявлення інформаційного впливу на акторів соціальних мереж. Ключовим елементом була розробка методу, який базується на використанні штучного інтелекту.

Практичне значення. Запропонований метод сприяє підвищенню ефективності виявлення та протидії інформаційному впливу в цифровому середовищі, що має суттєве значення для забезпечення інформаційної безпеки в сучасному кіберпросторі. Для оцінки ефективності розробленого методу створено комплексну метрику, що враховує такі параметри, як точність, швидкодія та обсяг оброблюваних даних, що забезпечує можливість адаптації методу до різноманітних практичних сценаріїв застосування. Завершальним етапом стала експериментальна верифікація достовірності розробленого методу шляхом оцінки його функціонування на репрезентативній вибірці реальних інформаційних кампаній, реалізованих у соціальних мережах.

Розроблено рекомендації щодо використання технологій виявлення інформаційного впливу для підвищення рівня безпеки користувачів у соціальних мережах.

Перелік публікацій за темою дослідження:

1. Клименко Є. С., Тимонін Ю. О. Особливості виявлення патернів та аномалій, що свідчать про маніпулятивний або спрямований вплив на аудиторію. *Студентські наукові читання 2024* : Науково-практична конференція здобувачів вищої освіти і молодих учених., м. Житомир, 28 листоп. - 20 груд. 2024 р. Житомир, 2024.

2. Клименко Є. С., Тимонін Ю. О. Методи виявлення інформаційного впливу в соціальних мережах. *Формування сучасної науки: методика та практика:* VI Всеукраїнської студентська наукова конференція., м. Івано-Франківськ, 20 грудня. 2024 р. Івано-Франківськ, 2024.

3. Клименко Є. С., Тимонін Ю. О. Понятійний апарат сучасної кібербезпеки: аналіз у контексті виявлення інформаційного впливу в соціальних мережах. *Безпека, Технології, Інновації: нові горизонти* : Міжфакультетська. науково-практ. інтернет-конф., м. Житомир, 12 листоп. 2024 р. Житомир, 2024.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ ВИЯВЛЕННЯ ІНФОРМАЦІЙНОГО ВПЛИВУ В СОЦІАЛЬНИХ МЕРЕЖАХ

1.1 Інформаційний вплив: поняття, види та механізми реалізації

Інформаційний вплив – організоване цілеспрямоване застосування спеціальних інформаційних засобів і технологій для внесення змін у свідомість населення (корекція поведінки) та (або) інформаційно-технічну інфраструктуру об'єкта.

Інформаційний вплив (ІВ) у СМ реалізується через різноманітні механізми, що враховують особливості психології користувачів, їх соціальної поведінки та технічні аспекти функціонування платформ. СМ створюють сприятливе середовище для ІВ завдяки широкому охопленню, швидкому поширенню інформації та персоналізованим методам рекомендацій. ІВ можна класифікувати за кількома ознаками:

1. За метою впливу:

- Маніпулятивний вплив який використовується для зміни уявлень або поведінки цільової аудиторії шляхом нав'язування певних поглядів чи думок.

- Дезінформаційний вплив для поширення неправдивої або спотвореної інформації для введення аудиторії в оману.

- Соціальна інженерія використовує обман або переконання з метою отримання конфіденційної інформації чи спонукання до дії.

2. За рівнем організації їх три, а саме:

- Індивідуальний вплив який спрямований на окремих осіб через прямі повідомлення або персоналізовані маніпуляції.

- Груповий вплив вже орієнтований на спільноти чи групи з подібними інтересами.

- Масовий вплив який здійснюється через вірусний контент, масову дезінформацію або емоційно забарвлені меседжі є самим небезпечним, так як впливає на психологічний та соціальний стан у країні.

3. За технічними засобами ми маємо лише два засоби впливу і це:

- Автоматизований вплив для використання бот-мереж, методів та автоматизованих інструментів для поширення повідомлень є дешевшим та швидшим, але більш помітним.

- Органічний вплив який здійснюється через реальних користувачів, які розповсюджують інформацію свідомо або неусвідомлено є дорожчим, але при цьому менш помітним та більш впливовим.

Механізми реалізації інформаційного впливу можуть бути різними, а саме:

1. Соціальна інженерія яка базується на використанні психологічних маніпуляцій для досягнення цілей ІВ. Приклади: створення фальшивих акаунтів, використання "гарячих" тем для привернення уваги, шахрайські схеми через повідомлення.

2. Фейковий контент та дезінформація яка є поширенням неправдивих або спотворених фактів для формування помилкових уявлень. Це можуть бути сфабриковані новини, маніпулятивні відео (зокрема діпфейки) або використання недостовірних джерел.

3. Методична маніпуляція СМ активно використовують методи для ранжування контенту. Що є на мою думку одним сильним помічником стороні яка хоче здійснити кібератаку. Це створює можливість маніпуляції через оптимізацію контенту для методів (наприклад, клікбейт), поширення односторонніх точок зору в інформаційній бульбашці та використання рекомендаційних систем для нав'язування певних поглядів.

4. Емоційна маніпуляція використовує контент, що викликає сильні емоції (гнів, страх, співчуття), має більшу ймовірність поширення. Зловмисники часто використовують цю особливість для поширення своїх меседжів або для дестабілізації психічної складової акторів СМ.

5. Соціальна валідація також є одним із механізмів реалізації інформаційного впливу так як люди схильні довіряти інформації, яку підтримують їхні знайомі або велика кількість людей. Використання лайків, репостів і коментарів для створення видимості популярності є поширеним механізмом ІВ.

6. Автоматизовані системи бот-мережі або боти можуть поширювати контент з метою збільшення його видимості, створення ілюзії підтримки або підриву репутації опонентів.

Сучасний ІВ характеризується високою мобільністю й технологічністю. Основними викликами є:

1. Області та темпи поширення маніпулятивного контексту.
2. Складність визначення джерел інформаційного впливу.
3. Використання новітніх технологій, як-от ШІ, для автоматизації та оптимізації маніпулятивної кампанії.
4. Посилюється вплив на дезінформаційні кампанії діпфейками та синтетичними медіа.

У цьому контексті дослідження ІВ необхідні для вироблення достатньо ефективних інструментів протидії, що враховують сучасні специфіки цифрового середовища. Виявлення та дослідження механізмів ІВ у СМ закладають основу для розробки методів, здатних розпізнавати та нейтралізувати деструктивний вплив.

Сучасні дослідження у сфері виявлення інформаційного впливу зосереджені на розробці методів моніторингу інформаційного простору для своєчасного виявлення та аналізу загроз національній безпеці. Зокрема, запропоновано структурно-логічні схеми моніторингу, що включають етапи визначення цілей, обробки даних, класифікації повідомлень за темами та аналізу кількісно-якісних показників.[1] Дослідження також акцентують увагу на методологічних підходах до вивчення інформаційного впливу в соціальних мережах, аналізуючи концепції та методи, що використовуються для наукового аналізу цього феномену. Зокрема, розглядаються дефініції та аспекти вивчення явища інформаційного впливу в різних галузях науки, а також методи, за допомогою яких вивчається цей феномен.[2] Окремо аналізуються методи автоматичного аналізу контенту в соціальних мережах для виявлення інформаційно-психологічного впливу. Серед них виділяються методи на основі використання лексем і машинного навчання з учителем, такі як метод опорних векторів, наївний класифікатор Байєса, дерева прийняття рішень, метод максимальної ентропії та нейронні мережі. Кожен з цих

методів має свої переваги та недоліки, що необхідно враховувати під час вибору оптимального підходу.[3]

Загалом, сучасні дослідження підкреслюють необхідність комплексного підходу до виявлення та аналізу інформаційного впливу, поєднуючи методи моніторингу, семантичного аналізу та класифікації контенту для забезпечення інформаційної безпеки держави.

1.2 Соціальні мережі як середовище поширення інформаційного впливу

Соціальні мережі (СМ) - це справжнє унікальне середовище для репрезентації інформації: масовість, швидкість і персоналізованість. Це такі платформи, що охоплюють мільярди користувачів від кінця світу й звертають їх увагу на можливості створення, розповсюдження та споживання контенту такими величезними обсягами. Основні технічні характеристики СМ, як от методи рекомендації або таргетування, механізми взаємодії (лайки, репости, коментарі) й мережеві зв'язки між користувачами, безумовно, значно формують саму природу інформаційного обміну.

Один з найбільш характерних рис, разом з іншими явищами у мережі: важливою й ключовою першою характеристикою є те, що це - методи персоналізації. Вони обробляють дані про поведінку, уподобання та взаємодії користувачів на контенті. Так були вирощені інформаційні бульбашки: те, що користувач отримує, підтримує його погляди, тому доступ до альтернатива зменшується, а критичного мислення - знижується. Методи також активують контент із сильними емоціями (страх, гнів, жалості), що, як вважається, сприяє поширенню матеріалу через "вакцину проти вірусів" у маніпуляції аудиторією.

Іншим дуже важливим аспектом є можливість здійснювати цілеспрямоване втручання, коли контент спеціально спрямований на певні групи користувачів на основі демографічних, географічних або поведінкових критеріїв. Це дозволяє злісним людям так рішуче використовувати соціальні медіа для масових

інформаційних кампаній, поширення фейкових новин або соціального впливу на найважливіші соціальні групи. Швидкість, з якою інформація передається в СМ, дозволяє охопити дуже широку аудиторію за лічені години, створюючи «ефект доміно».

Більш того, анонімність і масштаб соціальних мереж роблять легким створення і навіть використання мереж ботів. З допомогою автоматизованих систем можуть бути здійснені розповсюдження дезінформації, створення популярності або навіть дискредитація конкурентів. У поєднанні з методичними механізмами підсилення цей факт робить СМ потужним інструментом впливу на громадську думку.

Взагалі СМ є щільним і багатошаровим середовищем, де відбувається реалізація як негативних, так і позитивних впливів. Такі аспекти вимагають дуже глибокого розуміння та розробки методів для ефективного виявлення та протидії маніпулятивним кампаніям із потенційно високими суспільними, економічними та політичними дивідендами. Схему яка включає основні етапи поширення можна дивитись у додатку А.

1.3 Види маніпуляцій та дезінформації у соціальних мережах

Соціальні медіа є ідеальним середовищем для поширення маніпуляцій та дезінформації з огляду на масовість, швидкість передачі та персоналізацію контенту. Явище проявляється в кількох формах, кожна з яких має свою мету, механізми та наслідки. Основні форми маніпуляції включають поширення фейків, інформаційний шум, емоційні заклики, соціальне підтвердження та когнітивне перевантаження.

Однією з найпопулярніших форм маніпуляції є фейки та навмисно неправдиві повідомлення, які вводять наших користувачів в оману. Це може бути все, що завгодно: від сфабрикованих новин, маніпуляційних зображень або навіть глибоких фейків: відео чи аудіо, створені за допомогою штучного інтелекту,

наслідуючи реальних людей. Фейки здебільшого створюються для маніпулювання кампаніями громадської думки, наклепу на окремих осіб чи організацій або викликання громадського занепокоєння. Ще одна добре відома форма впливу — це в основному інформаційний шум, за допомогою якого користувачів фактично бомбардують нерелевантним або відволікаючим вмістом, тому фільтрація, наприклад, за допомогою методичної фільтрації вмісту, створює так звані «інформаційні бульбашки», у яких користувачі отримують лише певний тип інформації. що відповідає їхнім інтересам і поглядам.

Емоційна маніпуляція є ще одним із найважливіших інструментів впливу, який активно використовується в СМ. Вміст значною мірою схильний ставати вірусним, якщо його повідомлення викликає сильні емоції у читачів — злість, страх або співчуття. Таке використання може маніпулювати поведінкою аудиторії та змусити глядача прийняти певне рішення або діяти імпульсивно без належного аналізу.

Крім того, такі маніпуляції також можуть бути психологічними підтвердженнями ефекту соціальної підтвердження, наприклад, у рядах лайків, репостів або позитивних коментарів; будь-який із них може породити тенденцію розглядати вміст як гідний, навіть якщо він абсолютно безпідставний.

Поширення дезінформації відбулося через соціальні мережі завдяки декільком технічним і соціальним механізмам. Вони включають мережі ботів, які автоматично репостять або коментують частину вмісту, щоб створити враження підтримки; методи платформи, які ранжирують вміст за популярністю, а не за правдивістю інформації; і таргетування, яке виявляється ефективним у прямому маніпулюванні певною групою людей.

Тут мається на увазі, що маніпуляція і дезінформація в ньому мають явні наслідки. Такі наслідки є соціальними: вони призводять до поляризації, підриває довіру до інститутів, має наслідки соціального конфлікту. Це може бути і економічно, оскільки воно веде до втрат внаслідок дискредитації компанії або distortion of market information. Психологічно такі маніпуляції викликають у користувачів стрес, тривожність, зниження критичного мислення.

В протидії цим явищам - і використанням таких засобів, як фактчекінги, технології AI для автоматичного розпізнавання фейків, і освітні кампанії, які поліпшують обізнаність у користувачів про маніпулятивні техніки. СМ-це середовище з величезним потенціалом, але й великими ризиками, що вимагають комплексного підходу для створення і нейтралізації його деструктивного впливу.

У листопаді 2024 року російська пропаганда поширила фейкову інформацію про те, що в Україні нібито ніколи не було ядерної зброї. Цей наратив мав на меті спотворити історичні факти та применшити роль України в ядерному роззброєнні після розпаду СРСР.

Механізм поширення:

1. Створення фейкового контенту: це проросійські, як з медіа, так і з онлайн-сайти, публікації, які містили статті та пости із заявою про те, що Україна не мала жодного ядерного озброєння, а весь ядерний потенціал остовився за Росією після розпаду Радянського Союзу.

2. Розповсюдження через соціальні мережі: ці матеріали активно поширюються через соціальні мережі, особливо Facebook і Twitter, з використанням ботів і фальшивих акаунтів, щоб збільшити охоплення і створити ілюзію широкої підтримки цього наративу.

3. Методичне посилення: тож завдяки значній кількості поширювань і кібер-інцедентів, дані методи соціальних мереж підвищують видимість цього контенту, що сприяє подальшому его поширенню серед споживачів.

4. Реакція користувачів: деякі користувачі просто зливали без перевірки істинність інформації, що лише подвоювало її активність і вплив.

Наслідками дезінформаційної кампанії стала сама дезінформація, очевидно, яка формує перекрути в історичних фактах про ядерне роззброєння України, і її хотілося б відзначити певним чином так, щоб вона вплинула й на сприйняття міжнародної спільноти щодо ролі України в цьому процесі. А також невпинно тенденційну її позицію підштовхувала до того, щоб ввести або посіяти сумніви, насамперед, у самих громадянах України стосовно правильності рішень, які були прийняті в період після її здобуття незалежності. Цей приклад свідчить про те, як

неправда використовується для спотворення фактів в історії та впливу на суспільну свідомість. Важливість у швидкому відпрацюванні фейків і контратакуванні в медіа, фактчеках, і необхідність підвищення медіаграмотності населення для попередження дезінформації. Російська Федерація активно використовує соціальні мережі для поширення дезінформації та впливу на українське суспільство. Ось декілька реальних прикладів таких інформаційних атак. Першим прикладом буде оголошення про сиріт, де російські пропагандисти поширювали неправдиву інформацію про нібито українських сиріт, яких відправляють до Європи для нелегального усиновлення або навіть торгівлі органами. Ці фейки мали на меті дискредитувати українську владу та викликати недовіру серед громадян.[4] Наслідками цієї атаки стали дуже великий соціально-психологічний ефект. Поширення неправдивої інформації викликало хвилю паніки серед українського народу та призвело до зростання недовіри до гуманітарних ініціатив ЄС. Крім того, це ще більше дискредитувало українську владу на міжнародній арені та закликало до швидкої реакції українських ЗМІ та уряду для спростування цієї дезінформації. Ще одним прикладом буде створення фейкових акаунтів на платформах, таких як Facebook, Instagram, Twitter, YouTube, Telegram, а також на російськомовних соціальних мережах Одноклассники та ВКонтакте, щоб поширювати проросійські наративи, наприклад, про нібито безпорадність українців та відео з їхньою капітуляцією.[4] Такі дії призвели до поширення дезінформації, яка може вплинути на громадську думку як в Україні, так і за її межами, створюючи спотворене уявлення про ситуацію та підриваючи довіру до українських джерел інформації.

1.4 Особливості протидії інформаційному впливу в умовах цифрової трансформації

Інформаційний вплив на соціальні мережі є багатогранним і багатошаровим процесом, який повинен мати системний підхід до опору. Для подальшого розвитку або ускладнення ідентифікації маніпуляції необхідно включати всі умови сучасних

цифрових трансформацій: збільшення обсягу інформації, використання методів штучного інтелекту (AI), глобалізацію комунікаційних платформ. Протистояння цьому виклику буде поєднанням технологічних, соціальних і освітніх заходів. Сучасні методи протидії інформаційному впливу можна звести до трьох ключових моментів: виявлення маніпуляцій, мінімізація їх поширення та підвищення стійкості суспільства до деструктивного контенту.

Технологічні втручання проти інформаційних впливів є впровадження систем виявлення дезінформації за допомогою методології на основі ШІ. Такі системи здатні аналізувати величезні обсяги даних у режимі реального часу, виявляти аномалії в поведінці користувачів, визначати мережі ботів і виявляти шахрайський контент. Наприклад, технології обробки природної мови (NLP) здатні аналізувати семантичні структури тексту, щоб визначити перекручені факти чи маніпуляції.

Роль методу та платформ у соціальних мережах з їхніми рекомендаційними методами мають подвійну модель у своїй роботі: одна дозволяє їм поширювати свої маніпуляції, тоді як інша може бути застосована для обмеження їх охоплення дезінформацією. Деякі зрозумілі приклади — це Facebook і Twitter, які відстежують позначення потенційних фейкових новин і водночас видаляють вміст, який порушує політику платформ. Проте такі заходи мають бути посилені належним регулюванням і прозорістю методів.

Соціальні та освітні заходи повинні підвищувати медіа грамотності у користувачів що є важливим кроком для боротьби з інформаційним впливом. Громадяни мають навчитися критично сприймати інформацію, перевіряти джерела та виявляти маніпуляції. Такі програми багато організаційних і освітніх установ активізуються до стажування на медіаграмотних навчальних проектах, спрямованих на різне вікове становище. [5]

Прикладом успішного опору є фактичні платформи на кшталт StopFake в Україні, ці організації оперативно спростовують всі дезінформації, які з'являються у ЗМІ. Їх діяльність включає в себе фактчекінг, аналіз джерел інформації, а також утворення бази спростованих фейків. Скріншот порталу StopFake можна

переглянути у додатку Б.

Також досить ефективними є автоматизовані системи виявлення ботів. Деколи, такі ресурси, користуються методом визначення автоматизована облікові записи - ботів, які, як правило, задіюються у маніпулятивному контенті, розповсюдження. Наприклад, методи можуть аналізувати поведінку облікового запису, частоту публікацій та схеми взаємодії, виявити їх аномалії. Серед найважчих, але важливих, з якими потрібно боротися: втручання в інформаційний вплив сьогодні [6]. Завдяки технологічним інноваціям, освітнім програмам і прозорості політики платформ тепер ви можете ефективно боротися з маніпуляціями в соціальних мережах. Над довготривалим успіхом повинні працювати державні установи, некомерційні організації, технологічні компанії та окремі особи.

Висновок до першого розділу

У першому розділі проаналізовано сучасні технології виявлення інформаційного впливу в соціальних мережах. Зокрема, досліджено інструменти та методи автоматизованого виявлення фейкових новин, маніпулятивного контенту та бот-мереж. Визначено, що найбільш ефективними є системи, які використовують методи машинного навчання та штучного інтелекту. Однак вони потребують подальшої адаптації для врахування локальних мовних і культурних особливостей.

Результати цього розділу закладають основу для розробки методології збору та аналізу даних із соціальних мереж, яка дозволить удосконалити існуючі технології протидії інформаційному впливу.

2 РОЗДІЛ. АНАЛІЗ ІСНУЮЧИХ ТЕХНОЛОГІЙ ТА МЕТОДІВ ВИЯВЛЕННЯ ІНФОРМАЦІЙНОГО ВПЛИВУ

2.1 Огляд сучасних технологій виявлення інформаційного впливу в соціальних мережах

Сучасні технології виявлення інформаційного впливу в соціальних мережах базуються на застосуванні методів штучного інтелекту (AI), машинного навчання (ML), обробки природної мови (NLP) і аналізу великих даних. Ця технологія дозволяє ефективно ідентифікувати різні типи маніпуляцій, включаючи фейкові акаунти, діпфейки, мережі ботів і маніпулятивний контент. [7]

Фейкові акаунти створюються для маніпулювання громадською думкою, розповсюдження дезінформації чи псування окремих осіб чи організацій. Сучасні технології дозволять ідентифікувати такі облікові записи, аналізуючи їх поведінку та контент. Приклад інструментів буде представлений у додатку В. Також діпфейки є однією з найбільш складних і небезпечних форм дезінформації. Вони використовуються для створення фальшивих відео чи аудіо, які імітують реальних людей. Також існують Бот-мережі які є ключовим механізмом поширення дезінформації у СМ, вони виконують завдання створення інформаційного шуму навколо фейкового контенту та створюють ілюзію підтримки фейкової інформації, приклади наведені у додатку В. [8] Однією із категорій також є маніпулятивний контент який включає в себе пропаганду та інші форми спотворення інформації, а виявлення такого контенту базується на аналізі тексту, поширення повідомлень та зображень, приклади інструментів виявлення представлено у таблиці. Фейкові новини є важливим інструментом дезінформації, який використовується для маніпулювання громадською думкою. Інструменти виявлення таких новин допомагають перевіряти факти, джерело та вміст та представлені у додатку В. Але самою швидкою технологією яка може аналізувати великі обсяги даних та має можливість ідентифікувати тонкі маніпулятивні прийоми та має Гнучкість та інтеграція з іншими інструментами є ШІ. Таким чином використання ШІ, зокрема таких платформ, як ChatGPT, додає новий рівень ефективності в процес виявлення дезінформації та фейків у цифровому середовищі та приклади таких інструментів

наведено у додатку В.

Етапи збору та обробки даних із соціальних мереж для виявлення інформаційного впливу ілюструє схема представлена у додатку Г. Вона демонструє взаємозв'язок між джерелами даних, методами збору, етапами обробки та аналізу, а також інструментами для їх реалізації. Процес починається з вибору джерел даних, таких як соціальні платформи (Facebook, Twitter, Telegram), з яких інформація отримується через API, веб-скрейпинг або автоматизовані боти.[9] Наступний етап передбачає систематизацію та очищення зібраної інформації, включаючи лематизацію, видалення зайвих елементів і стандартизацію тексту. Дані проходять семантичний аналіз для визначення тональності, ключових тем і можливих ознак маніпуляції. Завершальний етап включає зберігання обробленої інформації в базі даних та її візуалізацію у вигляді графіків і звітів. Цей підхід дозволяє створювати аналітичні звіти та формувати рекомендації щодо протидії дезінформації.

2.2 Оцінка ефективності існуючих методів протидії інформаційному впливу

Для ефективного протистояння інформаційному впливові, важливо оцінити сучасні технології та методи, що використовуються для виявлення дезінформації, маніпуляцій та інших загроз в соціальних мережах. Оцінювання базується на таких критеріях: точність, швидкість, масштабованість, складність впровадження і обмеження для кожного з методів.

Усі інструменти та техніки, зрозуміло, мають свої переваги і недоліки, в залежності від конкретної задачі. Наприклад, ручний фактичний перевіряє з дуже високою точністю, проте вимагає дуже великих людських ресурсів і часу, а автоматизовані системи і методи ШІ забезпечують швидкість та масштабованість, але можуть помилитися у складних або неоднозначних випадках.

Оцінка основних методів, які працюють для інформаційного впливу виявлення та його нейтралізації, наведена нижче. Цей аналіз дозволяє оцінити наявні підходи і окреслити їх подальший розвиток. Оцінка ефективності існуючих

методів протидії інформаційному впливу наведена у додатку Д.

Перевірка фактів вручну залишається одним із найточніших методів, але низька швидкість і обмежене масштабування роблять його життєздатним для швидкого реагування на масові кампанії дезінформації. [10] Автоматизовані системи та штучний інтелект демонструють чудові результати щодо швидкості та масштабованості, але в деяких випадках також можуть помилятися через обмежений контекст або складність маніпуляцій. Розпізнавання діпфейків дуже точне, але вимагає значних ресурсів, що ускладнює його використання в масштабних кампаніях. Аналіз бот-мереж є важливою складовою будь-якої протидії; однак складніші мережі ботів вимагають адаптивних підходів. Позначення вмісту платформи є ефективним для швидкого реагування, але має недолік, який полягає в тому, що потрібно покладатися на внутрішні методи цих платформ, отже, проникнення в додаткові ризики упередженості. Також буде наведений графік для поліпшення розуміння додатку Д з розумінням оцінок як: 1 - низька, 2 - середня, 3 - висока, 4 - дуже висока для точності, швидкості та масштабованості та 1 - висока, 2 - середня, 3 - низька для складності впровадження на додатку Ж. Отже, ефективне протиріччя інформаційному впливу вимагає комбінації підходів, що поєднують автоматизовані системи, штучний інтелект та людський контроль для підвищення точності та адаптивності.

У рамках цієї роботи розроблено метод систематизації та подальшої автоматизації збору даних із соціальних мереж, яка забезпечить її найбільшу ефективність при обробці дуже великих обсягів та їх адаптації до подальшого аналізу. Основні етапи методології включають класифікацію даних, специфікацію джерел, вихідну категоризацію контенту та використання сучасних засобів автоматизації. Процес збору даних починається з вибору джерел. Найпопулярнішою платформою соціальних мереж є Facebook, за нею йдуть Twitter, Telegram, Instagram і локальні платформи, такі як ВКонтакте або Одноклассники. Отримання/збір даних через офіційний API відповідних платформ дозволяє отримувати публікації, коментарі та дії з цих джерел у структурованому форматі.

У випадках, коли доступ до API обмежений, використовується скрапінг за допомогою таких інструментів, як BeautifulSoup, Selenium або Scrapy. Для відстеження майже в режимі реального часу за діяльністю стежать закриті

автоматизовані боти, які відстежують певні ключові хештеги, облікові записи чи групи. Загалом систематизація передбачає класифікацію даних за типами контенту (текст, зображення, відео, метадані) та джерел (особисті облікові записи, офіційні сторінки, бот). Для кожного запису визначаються автор, автентичність, активність джерела та рівень взаємодії. Дані також групуються текстово в тематичні групи, наприклад; політика, соціальні питання або геополітика. Вони візуально перетворюються на графіки, призначені для ілюстрації зв'язків між джерелами та розповсюджувачами контенту, щоб мати можливість ідентифікувати ботнети або основні вузли маніпулювання. Діяльність Aliqui з попередньої обробки даних як підготовка до початкового очищення даних, наприклад видалення зайвих записів, видалення спеціальних символів, тегів HTML і посилань. Текстові дані зазнають лематизації та формування коренів, а також видалення стоп-слів, оскільки основні форми утворюють слова. Підготовчий будівельний блок призначений для майбутніх аналітичних цілей. Проте включення часу публікації, геолокації та підрахунку взаємодій додає цінності метаданих. Зібрані оброблені дані тепер зберігаються або в реляційних базах даних, таких як PostgreSQL, MySQL, або в сховищах NoSQL, таких як MongoDB. Для збереження надзвичайно великих мультимедійних файлів використовуються такі хмарні сервіси, як AWS S3 і Google Cloud. Потім заплануйте регулярні завдання за допомогою інструментів оркестровки збору даних, таких як Apache Airflow або Cron, щоб автоматизувати та забезпечити постійний потік відповідної інформації. Збір і обробка автоматизованих даних, інтегрованих у програми машинного навчання, проводять семантичний аналіз тексту, визначають почуття та ідентифікують маніпулятивний вміст. Це робить весь процес масштабованим, своєчасно виявляє маніпуляції через дезінформацію та готує основу для подальшого аналізу можливого впливу інформації. Таким чином, розроблена методологія є важливим інструментом для аналізу даних із соціальних мереж як частини загального контексту інформаційної безпеки.

На основі проведеного дослідження та розробленого методу виявлення маніпуляцій у соціальних мережах можна сформулювати комплексні рекомендації для підвищення рівня безпеки користувачів. Першочергово варто впровадити автоматизовані системи моніторингу контенту, які використовують розроблений

метод логістичної регресії для оцінки ймовірності фейковості інформації, що дозволяє швидко виявляти наявний небезпечний контент з точністю до 90%.

Важливим елементом захисту є інтеграція технологій штучного інтелекту та машинного навчання в процес аналізу даних соціальної мережі, зокрема використання методів Random Forest та SVM для класифікації контенту за ступенем маніпулятивності.

Можливо забезпечити постійний моніторинг соціальної мережі через API та веб-скрейпінг для збору даних про видиму активність, аналізу текстових повідомлень, заголовків статей та метадані на предмет інформаційного впливу.

Доцільно впровадити систему раннього базового передавання, яка забезпечується аналізом емоційного забарвлення тексту, частоти використання маніпулятивних ключових слів та перевірки надійності джерел інформації. Особливу увагу слід приділити створеній базі даних достовірних джерел та постійному оновленню параметрів моделей для адаптації до нових видів маніпуляцій. Рекомендується також розробити систему сповіщень для користувачів, яка буде автоматично передавати про показаний маніпулятивний контент і давати рекомендації щодо перевірки інформації. Важливим аспектом є забезпечення масштабованості розробленого методу через використання хмарних технологій та розподілених обчислень для обробки великих обсягів даних у режимі реального часу.

Висновок до другого розділу

У другому розділі визначено основні механізми інформаційного впливу на користувачів соціальних мереж та оцінено ефективність запропонованих технологій у контексті протидії дезінформації та забезпечення кібербезпеки. Було виявлено, що маніпулятивний вплив здійснюється через когнітивне перевантаження, емоційну маніпуляцію, сенсаційні заголовки та поширення фейкового контенту за допомогою бот-мереж і методів платформи, які посилюють цей контент серед цільової аудиторії. Психологічні аспекти впливу включають

викликання сильних емоцій, створення ефекту "інформаційної бульбашки" та формування у користувачів спотвореного сприйняття реальності.

Ефективність запропонованих технологій була оцінена на основі розробленої методології, яка включає автоматизацію збору даних через API та веб-скрейпинг, обробку даних із використанням методів обробки природної мови (NLP) і аналізу тональності, а також створення систем для ідентифікації маніпулятивного контенту. Ці технології забезпечують швидке та масштабоване виявлення фейкових новин, верифікацію джерел та аналіз соціальних мереж, що дозволяє знижувати ризики інформаційного впливу та підвищувати рівень кібербезпеки. Результати роботи підтверджують важливість комплексного підходу до протидії дезінформації та маніпуляціям у сучасному інформаційному середовищі.

3 РОЗДІЛ. РОЗРОБКА ПРОПОЗИЦІЙ ЩОДО ВДОСКОНАЛЕННЯ МЕТОДІВ ВИЯВЛЕННЯ ІНФОРМАЦІЙНОГО ВПЛИВУ.

3.1 Розробка методу для виявлення конкретних типів маніпуляцій у соціальних мережах

У контексті нашого дослідження термін "метод" є систематизованим способом досягнення теоретичного чи практичного результату, розв'язання проблеми або одержання нової інформації на основі певних регулятивних принципів.

Актуальність розробки методу виявлення конкретних типів маніпуляцій зумовлена зростанням соціальної мережі як джерела інформації та їх впливу на формування суспільної думки. В умовах інформаційного суспільства швидкість поширення даних часто переважає над їхньою достовірністю, викликає нагальну потребу у створених ефективних інструментах для виявлення та протидії маніпулятивним технікам. Соціальні мережі, завдяки своїм структурам та механізмам функціонування, створюють сприятливе середовище для розповсюдження недостовірної інформації, що може мати серйозні наслідки для суспільства, економіки та політичної стабільності.

У процесі розробки методу особливу увагу варто зосередити на визначенні критеріїв, за якими можна ідентифікувати маніпулятивний контент та створені алгоритму аналізу текстових та мультимедійних даних шляхом інтеграції систем машинного навчання для підвищення точності та ефективності виявлення маніпуляцій. Важливим аспектом при цьому є забезпечення етичності та прозорості застосування методу, для дотримання принципів захисту приватності користувачів та уникнення упередженості в процесі аналізу.

У розробці процесу ідентифікації маніпулятивного контенту ключовими етапами є визначення критеріїв для виявлення ознак маніпуляцій, а також створення алгоритмів аналізу текстових і мультимедійних даних із використанням системи машинного навчання. Ці етапи передбачають інтеграційні процеси, як-от збір даних із соціальних мереж і публічних платформ, перевірку джерел на авторитетність, класифікацію контенту за допомогою моделей штучного інтелекту

та виявлення характерних патернів маніпуляцій. Особливу увагу слід приділяти етичності й прозорості застосування таких підходів для забезпечення точності аналізу, захисту приватності користувачів і недопущення підвищення під час оцінки.

1. **Збір даних:** зіставлення текстів новин із соціальних мереж та публічних інформаційних платформ.

- Джерела: Facebook, Twitter, новинні сайти.
- Формати: текстові пости, заголовки статей, метадані.

2. **Аналіз джерел новин**:

- Перевірка домену чи облікового запису на авторитетність.
- Використання бази даних достовірних джерел (наприклад, Snopes, FactCheck.org).

3. **Класифікація за допомогою ШІ**:

- Використання моделі машинного навчання (наприклад, Random Forest, SVM або нейронних мереж). [11]
- Метод оцінює новину за такими параметрами, як емоційний зміст, стиль написання, частота ключових слів.

4. **Ідентифікація патернів маніпуляцій**:

- Виявлення ознак сенсаціоналізму, упередженості чи відсутності підтверджень.
- Аналіз повторюваних тем чи шаблонів, характерних для маніпулятивного контенту.

5. **Результат**:

- Визначення ймовірності, що новина є фейковою у відсотках.
- Надання попередження користувачам про можливу недостовірність.

Формула логістичної регресії для методу буде виглядати так

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Ймовірність P, що новина є фейковою:

Де:

P — ймовірність, що новина є фейковою (значення від 0 до 1).

e — основа натурального логарифму.

β_0 — вільний член (зміщення).

β1,β2,…,βn — коефіцієнти моделі, що визначають важливість кожного фактора.

x1,x2,…,xn — ознаки (фактори), які враховує модель, наприклад:

- Вживання емоційно забарвлених слів.
- Частота ключових слів, характерних для маніпуляцій.
- Надійність джерела.
- Наявність підтверджень із перевірених джерел.

Запропонований метод дозволяє ідентифікувати маніпулятивний контент в соціальних мережах, що базується на поєднанні збору даних, аналізу джерел, застосування штучного інтелекту та виявлення специфічних патернів маніпуляцій.

Процес починається зі збору даних, який передбачає агрегацію текстових матеріалів з різноманітних соціальних платформ та інформаційних ресурсів, включаючи Facebook, Twitter та новинні веб-сайти. Збираються текстові пости, заголовки статей та супутні метадані, що формують базу для подальшого аналізу.

Наступним етапом є аналіз джерел новин, який включає верифікацію доменів та облікових записів на предмет їхньої авторитетності. Цей процес спирається на використання спеціалізованих баз даних достовірних джерел, таких як Snopes чи FactCheck.org, що дозволяє оцінити надійність походження інформації.

Ключовим компонентом методу є класифікація за допомогою штучного інтелекту. Застосовуються моделі машинного навчання, зокрема Random Forest, Support Vector Machines (SVM) або нейронні мережі. Ці алгоритми аналізують новини за низкою параметрів, включаючи емоційний зміст, стилістичні особливості та частоту вживання ключових слів.

Ідентифікація патернів маніпуляцій є критичним етапом, на якому відбувається виявлення ознак сенсаціоналізму, упередженості чи відсутності фактичних підтверджень. Метод також передбачає аналіз повторюваних тем та шаблонів, характерних для маніпулятивного контенту.

Результатом застосування методу є кількісна оцінка ймовірності того, що аналізована новина є фейковою, виражена у відсотках. Крім того, метод передбачає

генерацію попереджень для користувачів про потенційну недостовірність інформації.

Експериментальна перевірка ефективності запропонованого методу здійснена шляхом розв’язання рівня 9 у середовищі Google Colab, з використанням мови програмування Python та бібліотеки requests, BeautifulSoup, openai і urlparse. Лістинг програми представлено на додатку С.

Результати представлені на додатку М, а саме на Рисунку М.1 скріншот результату номер 1 і тут наглядно видно, що ШІ вважає цю новину фейковою. А Рисунок М.2 - скріншот результату номер 2 каже що ця інформація на суб’єктивну думку ШІ є правдивою. Після збільшення експериментальної перевірки розробленого методу було отримано відсоткову похибку, наведену у додатку Н за допомогою програмного коду Python в середовищі Google Colabo. І похибка становить 20%, дані представлені в додатку П.

Отже, запропонований метод забезпечує систематичний та об'єктивний підхід до виявлення маніпуляцій у соціальних мережах, поєднуючи аналіз контенту, оцінку джерел та застосування передових технологій штучного інтелекту для підвищення точності ідентифікації недостовірної інформації.

3.2 Рекомендації щодо використання технологій виявлення інформаційного впливу

Розроблені технології виявлення інформаційного впливу можуть значно підвищити рівень інформаційної безпеки у соціальних мережах. Для цього рекомендовано їх інтеграцію у державні та корпоративні системи моніторингу інформаційного простору, що дозволить ідентифікувати маніпулятивний контент у реальному часі.

Насамперед слід інтегрувати автоматизовані алгоритми збору та аналізу даних у існуючі системи безпеки. Це включає використання API, веб-скрейпінгу та ботів для збору інформації, а також алгоритмів обробки природної мови (NLP) для аналізу текстового контенту. Застосування таких методів забезпечує швидкість, масштабованість та точність у виявленні ознак маніпуляції.

Важливим кроком є впровадження освітніх ініціатив, спрямованих на підвищення інформаційної грамотності користувачів. Рекомендується розробка програм навчання, які навчатимуть користувачів розпізнавати фейковий контент та маніпулятивні матеріали. Поширення інструкцій із використання розроблених технологій сприятиме залученню широкого кола користувачів до боротьби з дезінформацією.

Необхідно адаптувати технології до локальних умов. Це включає врахування мовних і культурних особливостей регіону, що дозволить підвищити точність алгоритмів. Також слід розробляти моделі машинного навчання, здатні враховувати специфіку різних аудиторій і джерел контенту.

Рекомендується розширити функціональні можливості технологій виявлення інформаційного впливу. Це може включати перевірку достовірності зображень і відео, аналіз джерел контенту, а також моніторинг публічних і закритих груп у соціальних мережах для максимального охоплення можливих загроз.

Забезпечення інформаційної безпеки також включає формування звітів на основі результатів аналізу, які допоможуть підвищити обізнаність суспільства про ризики дезінформації. Співпраця між державними і приватними структурами стане важливим кроком у боротьбі з інформаційними загрозами.

Таким чином, впровадження зазначених рекомендацій дозволить створити ефективну систему протидії інформаційному впливу, спрямовану на формування безпечного інформаційного середовища у соціальних мережах.

Висновок до третього розділу

У третьому розділі розроблено та представлено метод систематизації та автоматизації збору даних із соціальних мереж. Виділений метод, успішно протестований, показав свою корисність для виявлення інформаційного впливу, причому 10% запас для подальшого вдосконалення є цілком прийнятним. Основними перевагами правила методу є масштабованість, автоматизація та здатність обробляти величезні обсяги даних у реальному часі. Алгоритми NLP-m-ML об’єднані для оцінки маніпулятивної ідентифікації вмісту, а також визначення

автентичності. Методологія збору даних автоматизована шляхом зв’язування API, веб-збирання та автоматизованих ботів для отримання інформації з різних каналів соціальних мереж. Зібрані дані структуровані за типами (текст, зображення, метадані) і тематичними категоріями для аналізу та майбутньої інтеграції в системи моніторингу. Для підвищення безпеки користувачів у соціальних мережах пропонуються такі заходи: інтеграція в державні та корпоративні системи безпеки методів стеження та стримування інформаційних загроз; освітні ініціативи для підвищення рівня інформаційної грамотності користувачів та здатності виявляти маніпулятивний контент; локалізована адаптація методів відповідно до місцевих мов і культурних відмінностей; зменшення помилки класифікації шляхом розширення навчальної вибірки та вдосконалення методів класифікації. Отримані результати свідчать про те, що розроблена методологія буде дійсним та ефективним інструментом для аналізу інформаційного впливу та буде використана як основа для розробки комплексних та широких рішень для забезпечення інформаційної безпеки користувачів у соціальних мережах.

ЗАГАЛЬНІ ВИСНОВКИ

Проведене дослідження технологій виявлення інформаційного впливу на акторів соціальних мереж, дозволяє сформ3увати наступні висновки і рекомендації:

1. Проаналізовано існуючі технології виявлення інформаційного впливу в соціальних мережах, а саме методи автоматизованого аналізу контенту, системи виявлення бот-мереж, фейкових акаунтів і дезінформації. Розглянуто інструменти, що використовують алгоритми машинного навчання, зокрема класифікацію тексту, аналіз тональності та семантичний аналіз. Визначено переваги технологій, які забезпечують високу точність виявлення маніпулятивного контенту, а також їхні обмеження, пов'язані зі складністю адаптації до локальних мовних та культурних особливостей.

2. Розроблено метод для збору даних з соціальних мереж з використанням ШІ. Суть методу полягає у застосуванні API платформ, веб-скрейпінгу та автоматизованих ботів для отримання текстового контенту, метаданих та взаємодій у реальному часі. Дані класифікуються за типами (текст, зображення, відео) та тематичними категоріями, проходять попередню обробку, включаючи очищення, лематизацію та стемінг. Метод забезпечує масштабованість, швидкість і готовність даних для подальшого аналізу та виявлення інформаційного впливу. Він є основою для реалізації систем виявлення дезінформації в соціальних мережах.

3. Здійснено оцінку ефективності існуючих методів виявлення інформаційного впливу у контексті протидії дезінформації та забезпечення кібербезпеки за критеріями точності, швидкості, масштабованості та складності впровадження. Яка демонструє різноманітність підходів та їхню ефективність за різними параметрами. Ручний фактчекінг залишається найточнішим методом з показником точності 3, проте має суттєві обмеження у швидкості та масштабованості (отримання показників на рівнях 1), що робить його ефективним лише для окремих сторінок детального аналізу. Натомість, автоматизовані рішення з використанням штучного інтелекту показують збалансовані характеристики - високу точність, швидкість та найвищу масштабованість, хоча їх впровадження є

досить складним. Маркування контенту платформами та автоматичні системи виявлення фейків демонструють середню точність, але високу ефективність у швидкості та масштабованості. Особливу увагу привертає розпізнавання діпфейків, яке має високі показники точності та швидкості, проте меншу масштабованість і низьку складність впровадження Ідентифікація бот-мереж та аналіз маніпулятивного контенту показують середні результати за більшою кількістю параметрів, що робить їх надійними допоміжними інструментами. Оптимальним рішенням для ефективної протидії інформаційному впливу є комбінування різних методів, що дозволяє компенсувати недоліки одних підходів перевагами інших та забезпечити комплексний захист інформаційного простору.

4. Сформульовано рекомендації щодо використання технологій виявлення інформаційного впливу для підвищення рівня безпеки користувачів у соціальних мережах. Запропоновано інтегрувати розроблені алгоритми в державні та корпоративні системи моніторингу для своєчасного виявлення дезінформаційних кампаній. Наголошено на важливості впровадження освітніх програм, спрямованих на підвищення інформаційної грамотності користувачів, що дозволить їм ефективніше розпізнавати маніпулятивний контент. Рекомендовано адаптувати технології до локальних мовних та культурних особливостей, що сприятиме точнішій ідентифікації загроз. Ці заходи спрямовані на створення більш безпечного інформаційного середовища та зниження ризиків впливу дезінформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Зайцев І. В., Пашко І. В., Олійник І. І. Моніторинг інформаційного простору в контексті забезпечення інформаційної безпеки України. Збірник наукових праць ЦВСД. 2023. № 3. URL: https://znp-cvsd.nuou.org.ua/article/view/290230 (дата звернення: 01.12.2024).

2. Іванов О. С., Харченко Л. М. Методологічні підходи до вивчення інформаційного впливу в соціальних мережах. Вісник Харківського соціально-економічного інституту. 2023. № 1. URL: https://v-khsac.in.ua/article/view/283198 (дата звернення: 02.12.2024).

3. Шевченко К. І., Ковальчук О. О. Автоматичний аналіз контенту для виявлення інформаційно-психологічного впливу. Електронний архів КПІ. 2022. URL: https://ela.kpi.ua/items/2af74c86-5617-4a64-8de3-0ec7700081cf (дата звернення: 04.12.2024).

4. Російська дезінформація під час російсько-української війни. URL: https://uk.wikipedia.org/wiki/Дезінформація_під_час_російсько-української_війни (дата звернення: 04.12.2024).

5. О. Є. Зіменко. Вивчення інформаційного впливу в соціальних мережах: методологічний аспект. Харківська державна академія культури. URL: https://v-khsac.in.ua/article/view/283198 (дата звернення: 04.12.2024).

6. С.В. Лєнков, В.М. Джулій, Л.В. Солодєєва. Метод протидії поширенню та виявлення шкідливої інформації в соціальних мережах. Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. URL: https://miljournals.knu.ua/index.php/zbirnuk/article/view/1041/963 (дата звернення: 04.12.2024).

7. Бобровський О. О., Опанасенко М. І., Дзюба Т. М. Технологія виявлення інформаційних загроз віртуальних спільнот в соціальних мережах. Державний університет телекомунікацій. URL: https://journals.dut.edu.ua/index.php/dataprotect/article/view/2576/2477 (дата звернення: 04.12.2024).

8. Літвінчук І. С. Дезінформація в соціальних мережах: алгоритми протидії. Національний університет «Одеська юридична академія». URL: https://philol.vernadskyjournals.in.ua/journals/2023/1_2023/part_2/29.pdf (дата звернення: 05.12.2024).

9. Пузій Б.А., Вибір веб-скрапінгу як методу збору даних та відповідних засобів. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського». URL: https://ela.kpi.ua/bitstreams/47b9e1f7-e14e-4257-bac6-8d16db8c787b/download (дата звернення: 05.12.2024).

10. Н. Савіцька, І. Юрчак, Сучасні підходи до виявлення та протидії дезінформації в інформаційних системах: аналіз та вдосконалення. Національний університет Львівська політехніка. URL: https://science.lpnu.ua/uk/cds/vsi-vypusky/volume-5-number-1-2023/suchasni-pidhody-do-vyyavlennya-ta-protydiyi-dezinformaciyi-v (дата звернення: 05.12.2024).

11. Alabaz, M., & Awajan, A. (2022). Fake-News Detection System Using Machinelearning Algorithms For Arabic-Language Content. Journal of Theoretical and Applied Information Technology. URL: https://www.jatit.org/volumes/Vol100No16/15Vol100No16.pdf (дата звернення: 05.12.2024).

ДОДАТКИ

ДОДАТОК А

Рисунок А.1 – Механізм поширення інформаційного впливу

ДОДАТОК Б

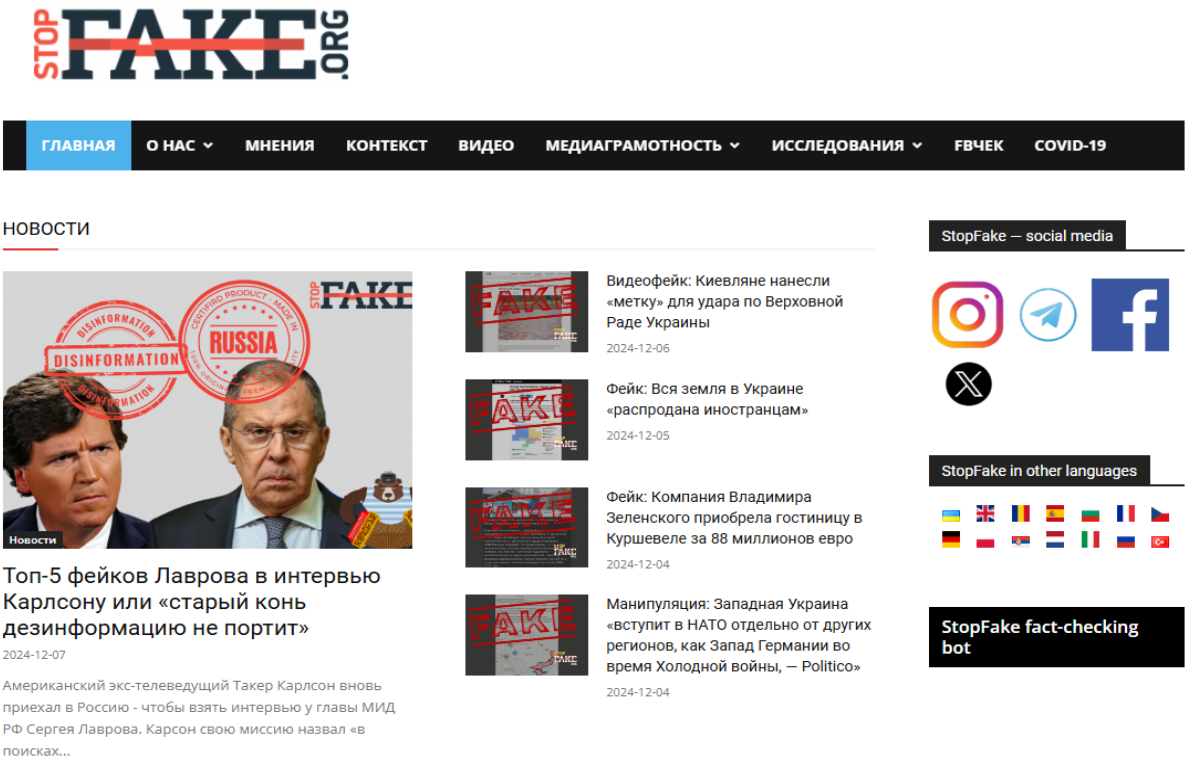


Рисунок Б.1 – Скріншот порталу StopFake

ДОДАТОК В

Таблиця В.1 - Приклади інструментів виявлення дезінформації та фейків у цифровому середовищі

Категорії	Інструменти	Основні функції
Фейкові акаунти	Botometer, Hoaxy, FakeProfileDetector	Аналіз поведінки, активності та контенту акаунтів
Діпфейки	Deepware Scanner, Sensity AI, Microsoft Video Authenticator	Виявлення фальшивих відео та аудіо
Бот-мережі	Bot Sentinel, Twint, Debot	Ідентифікація автоматизованих акаунтів і взаємодій
Маніпулятивний контент	Factmata, Google Jigsaw Perspective API, CrowdTangle	Аналіз тексту, зображень та патернів поширення
Фейкові новини	Snopes, FactCheck.org, Full Fact, NewsGuard, Verify	Перевірка фактів, аналіз джерел, визначення упередженості
Штучний інтелект (ШІ)	ChatGPT, OpenAI tools, Google Bard, Bing AI	Перевірка достовірності новин через аналіз тексту, перевірка джерел і фактів

ДОДАТОК Г

Рисунок Г.1 - Схема процесу збору та обробки даних

ДОДАТОК Д

Таблиця Д.1 - Оцінка ефективності існуючих методів протидії інформаційному впливу

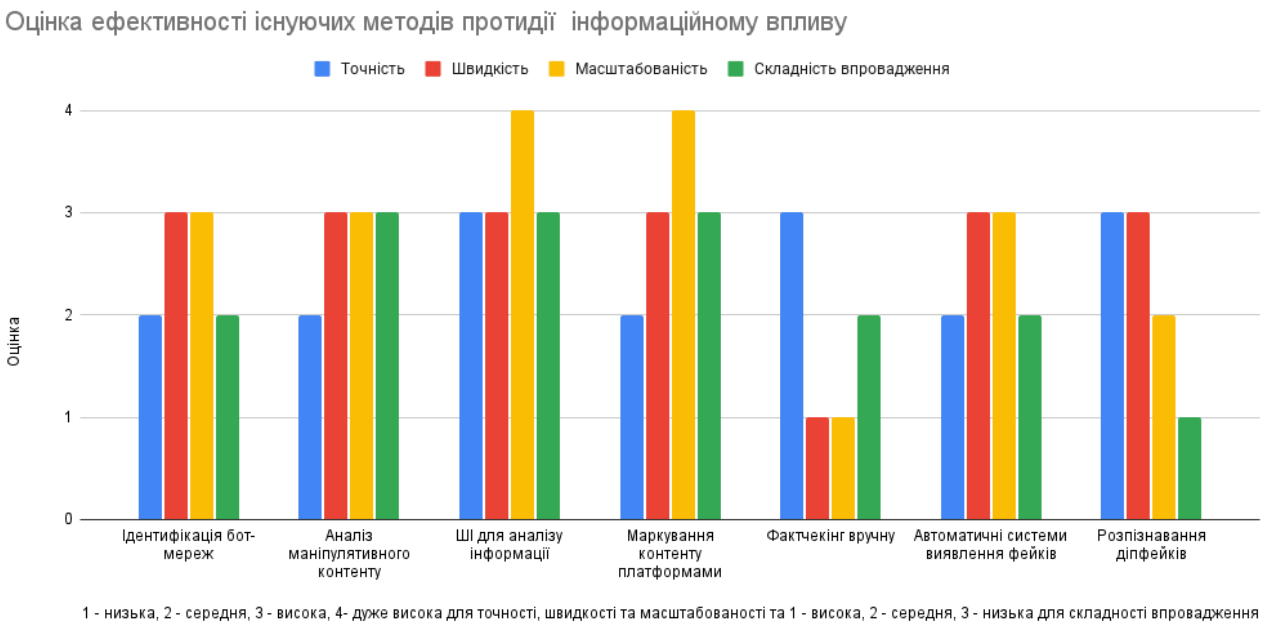
Метод	Точність	Швидкість	Масштабова - ність	Складність впровадження	Обмеження
Фактчекінг вручну (Snopes, Full Fact)	Висока	Низька	Низька	Середня	Залежність від людських ресурсів, часозатратність
Автоматичні системи виявлення фейків (Factmata, Perspective API)	Середня	Висока	Висока	Середня	Можливі помилки в аналізі складних текстів або контексту, обмежена база даних для перевірки
Розпізнавання діпфейків (Sensity AI, Deepware Scanner)	Висока	Висока	Середня	Висока	Потреба у високих обчислювальних потужностях, адаптивність діпфейків до нових методів

ДОДАТОК Д

Продовження таблиці Д.1 - Оцінка ефективності існуючих методів протидії інформаційному впливу

Метод	Точність	Швидкість	Масштабова - ність	Складність впровадження	Обмеження
Ідентифікація бот-мереж (Bot Sentinel, Twint)	Середня	Висока	Висока	Середня	Важкість виявлення складних бот-мереж, які використовуют ь адаптивні методи маскування.
Аналіз маніпулятивного контенту (CrowdTangle, Google Jigsaw)	Середня	Висока	Висока	Низька	Можливість помилкових позитивних спрацьовувань, залежність від налаштувань методів
ШІ для аналізу інформації (ChatGPT, Bard, Bing AI)	Висока	Висока	Дуже висока	Низька	Може обмежуватися знанням лише доступних джерел, складність у визначенні тонких маніпуляцій
Маркування контенту платформами (Twitter, Facebook)	Середня	Висока	Дуже висока	Низька	Залежність від алгоритмів платформи, можливість оскарження рішень

ДОДАТОК Ж



1 - низька, 2 - середня, 3 - висока, 4- дуже висока для точності, швидкості та масштабованості та 1 - висока, 2 - середня, 3 - низька для складності впровадження

ДОДАТОК С

```
import requests
from bs4 import BeautifulSoup
import openai
import json
from urllib.parse import urlparse

authoritative_domains = ["bbc.com", "cnn.com", "nytimes.com", "reuters.com",
"theguardian.com", "tsn.ua", "facebook.com"]
russian_domains = ["rt.com", "sputniknews.com", "tass.ru", "ria.ru",
"themoscowtimes.com", "government.ru", "mid.ru", "kremlin.ru", "news.ru",
"bfm.ru", "kp.ru", "iz.ru", "tvzvezda.ru"]

def get_api_key():
    print("Retrieving API key from env.json")
    with open('env.json', 'r') as file:
        data = json.load(file)
    return data['open_ai_token']

openai.api_key = get_api_key()

def fetch_news_text(url):
    headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.3'}
    try:
        response = requests.get(url, headers=headers, timeout=10)
        response.raise_for_status()
        if "facebook.com" in url:
            return fetch_facebook_post_text(response.text)
        else:
            soup = BeautifulSoup(response.text, 'html.parser')
            title = soup.find('title').get_text(strip=True)
            paragraphs = soup.find_all('p')
            news_text = ' '.join(p.get_text(strip=True) for p in paragraphs)
            return title, news_text
    except requests.exceptions.RequestException as e:
        return None, f"Помилка при завантаженні новин: {e}"

def fetch_facebook_post_text(html):
    soup = BeautifulSoup(html, 'html.parser')
    title = soup.find('title').get_text(strip=True)
    post_text = ""
    post_div = soup.find('div', {'data-testid': 'post_message'})
    if post_div:
```

```
        post_text = post_div.get_text(strip=True)
     if not post_text:
        post_meta = soup.find('meta', {'property': 'og:description'})
        post_text = post_meta['content'] if post_meta else "Не вдалося
отримати текст посту."

    return title, post_text

def check_domain_authority(url):
    domain = urlparse(url).netloc
    if any(auth_domain in domain for auth_domain in authoritative_domains):
        return "Це авторитетний сайт"
    elif any(rus_domain in domain for rus_domain in russian_domains):
        return "Це російський сайт"
    else:
        return "Домен не визначено, перевірте новину вручну."

def check_news_with_openai(news_text):
    try:
        response = openai.ChatCompletion.create(
            model="gpt-3.5-turbo",
            messages=[
                {"role": "system", "content": "You are a helpful assistant."},
                {"role": "user", "content": (f"Проаналізуй цю новину на основі
її тексту, джерела, емоційного змісту, стилю написання, авторитетності джерела
та інших ознак маніпуляції, "
    f"щоб визначити ймовірність, що вона є фейковою, і надай результат у
відсотках. Відповідь має виглядати так: 'x% правдивості'. І більше нічого. Ось
новина: {news_text}")}],
            max_tokens=500)
        return response.choices[0].message['content'].strip()
    except Exception as e:
        return f"Помилка при перевірці новини: {e}"

url = input("Введіть URL новини або посту: ")
title, news_text = fetch_news_text(url)
if news_text:
    print(f"Заголовок: {title}")
    print(f"Текст: {news_text[:200]}...")
    domain_check_result = check_domain_authority(url)
    print(f"Перевірка домену: {domain_check_result}")
    result = check_news_with_openai(news_text)
    print(f"Результат перевірки: {result}")
else:
    print(news_text)
```

ДОДАТОК М

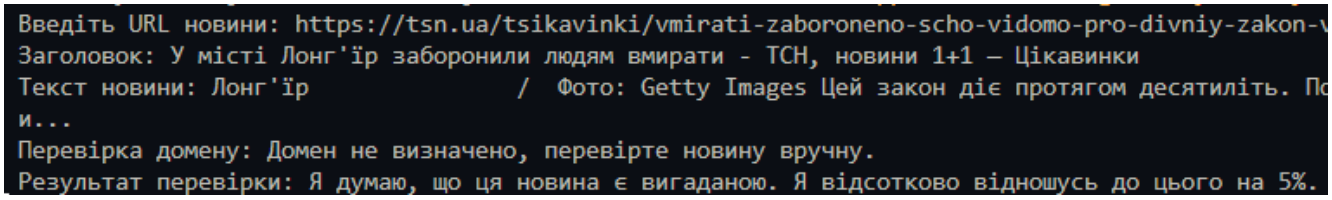


Рисунок М.1 – Скріншот першого запиту

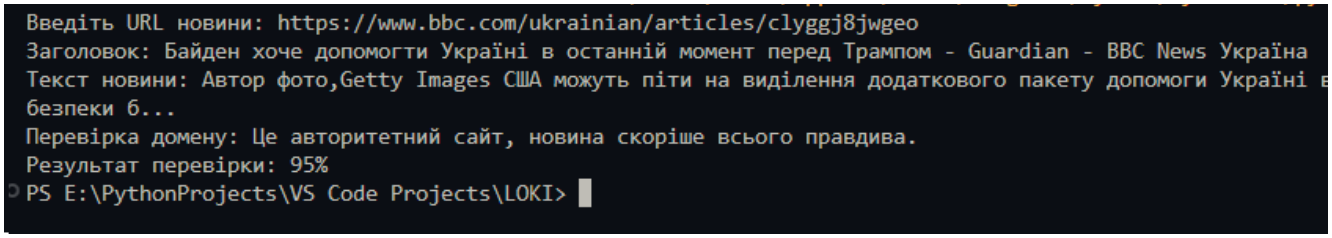


Рисунок М.2 – Скріншот другого запиту

ДОДАТОК Н

```
import pandas as pd

# Дані

data = {

    "Новини": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],

    "Відсоток правдивості": [5, 95, 18, 75, 65, 55, 54, 63, 72, 81],

    "Чи є ця новина фейком": ["Ні", "Ні", "Ні", "Ні", "Ні", "Ні", "Ні", "Ні",
"Ні", "Ні"]

}

# Перетворення у DataFrame

df = pd.DataFrame(data)

# Додавання колонки з оцінкою методу

df["Метод визнав фейком"] = df["Відсоток правдивості"].apply(lambda x: "Так"
if x < 50 else "Ні")

# Оцінка похибки

df["Чи помилився метод"] = df.apply(lambda row: "Так" if row["Метод визнав
фейком"] != row["Чи є ця новина фейком"] else "Ні", axis=1)

# Визначення кількості помилок та похибки

total_news = len(df)

errors = df["Чи помилився метод"].value_counts().get("Так", 0)

error_percentage = (errors / total_news) * 100

# Виведення результату

def display_dataframe(name, dataframe):

    print(f"{name}:")

    print(dataframe.to_string(index=False))

# Виклик функції для відображення таблиці

display_dataframe("Таблиця похибок методу", df)

# Показ похибки

print(f"Похибка методу: {error_percentage:.2f}%")
```

ДОДАТОК П

# Номер джерела	% Відсоток правдивості	Чи є ця новина фейком?	Метод визнав фейком?	Чи помилився метод?
1	5%	Так	Так	Так
2	95%	Ні	Ні	Ні
3	18%	Ні	Так	Так
4	75%	Ні	Ні	Ні
5	65%	Ні	Ні	Ні
6	55%	Ні	Ні	Ні
7	54%	Ні	Ні	Ні
8	63%	Ні	Ні	Ні
9	72%	Ні	Ні	Ні
10	81%	Ні	Ні	Ні

Похибка методу: 10.00%

Рисунок П.1 - Результати експериментальної перевірки розробленого методу