Харківський національний університет імені В. Н. Каразіна
ННІ "Комп’ютерних наук та штучного інтелекту"
Кафедра кібербезпеки інформаційних систем, мереж і технологій

«Допущено до захисту»
В. о. зав.кафедрою КІСМТ
Єсіна М. В._____
« » червня 2025 р.

Пояснювальна записка
до кваліфікаційної роботи бакалавра
спеціальність: 125 Кібербезпека

на тему: «Дослідження та аналіз можливостей штучного інтелекту та технології OSINT для виявлення шкідливого контенту в Telegram-каналах»

Оцінка « »
Голова ЕК
Мичуда Л.З.

Керівник: к.т.н, доцент
Єсіна М. В.
Рецензент: к.т.н., професор
Качко О. Г.
Виконавець: студент групи КБ41
Крижановський К. В.

Харків – 2025

РЕФЕРАТ

Пояснювальна записка містить 60 сторінки, 2 рисунки, 2 таблиці, 3 додатки, 51 джерело.

Метою кваліфікаційної роботи є аналіз існуючих інструментів для моніторингу та аналізу контенту в Telegram, а також дослідження використання штучного інтелекту та технологій OSINT для запобігання шкідливому контенту в інформаційному середовищі. Окрім, того завдання полягало у розробці інструменту, який інтегрує розглянуті технології, для автоматичного виявлення небезпечного контенту у Telegram.

Методи дослідження: аналіз наукових та технічних джерел щодо класифікації шкідливого контенту, практичного застосування методів обробки природної мови та глибокого навчання для роботи з даними, роль OSINT у виявленні шкідливого контенту, тестування та дослідження API Telegram.

У результаті дослідження проведено аналіз сучасних підходів до класифікації та ідентифікації шкідливого контенту, а також розроблено програмне рішення для автоматичного аналізу текстового контенту Telegram-каналів. Робота є новаторською тим, що інтегрує ШІ і OSINT для виявлення небезпек у середовищі Telegram, яке є нерегульованим на законодавчому рівні.

Запропоноване рішення може бути використане як основа для розробки систем автоматичного моніторингу соціальних платформ. Також результати роботи можуть стати корисними для фахівців, що займаються аналізом контенту.

Значущість роботи полягає у дослідженні можливостей штучного інтелекту та технологій OSINT для виявлення шкідливого контенту, а запропонований програмний підхід сприяє удосконаленню методів аналізу текстових даних у сфері кібербезпеки та інформаційного моніторингу.

Подальший розвиток дослідження може включати розширення функціональності системи для аналізу , а також адаптацію під інші платформи.

Ключові слова: ШТУЧНИЙ ІНТЕЛЕКТ, OSINT, TELEGRAM, ШКІДЛИВИЙ КОНТЕНТ, АНАЛІЗ ТЕКСТУ, МОДЕРАЦІЯ КОНТЕНТУ.

ABSTRACT

The explanatory note contains 60 pages, 2 figures, 2 tables, 3 appendix, and 51 sources.

The aim of the qualification work is to analyze existing tools for monitoring and analyzing content on Telegram, as well as to study the use of artificial intelligence and OSINT technologies to prevent harmful content in the information environment. In addition, the task was to develop a tool that integrates the technologies under consideration to automatically detect dangerous content on Telegram.

Research methods: analysis of scientific and technical sources on the classification of malicious content, practical application of natural language processing and deep learning methods for data processing, the role of OSINT in detecting malicious content, testing and research of Telegram API.

The research analyzes current approaches to classifying and identifying malicious content and develops a software solution for automatically analyzing textual content on Telegram channels. The work is innovative in that it integrates AI and OSINT to detect dangers in the Telegram environment, which is unregulated at the legislative level.

The proposed solution can be used as a basis for developing systems for automatic monitoring of social platforms. Also, the results of the work can be useful for specialists engaged in content analysis.

The significance of the work lies in the study of the possibilities of artificial intelligence and OSINT technologies for detecting malicious content, and the proposed software approach contributes to the improvement of methods for analyzing text data in the field of cybersecurity and information monitoring.

Further development of the research may include expanding the functionality of the system for analysis, as well as adaptation to other platforms.

Keywords: ARTIFICIAL INTELLIGENCE, OSINT, TELEGRAM, MALICIOUS CONTENT, TEXT ANALYSIS, CONTENT MODERATION.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ ТА СИМВОЛІВ 7

ВСТУП 9

1 ОГЛЯД ТА АНАЛІЗ ТЕОРЕТИЧНИХ ОСНОВ ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ В ІНФОРМАЦІЙНОМУ СЕРЕДОВИЩІ... 11

1.1 Визначення шкідливого контенту 11

1.2 Підхід до класифікації шкідливого контенту 11

1.3 Класифікація шкідливого контенту 13

1.3.1 Ненависть і переслідування 14

1.3.2 Самозаподіяння шкоди 21

1.3.3 Ідеологічна шкода 23

1.3.4 Використання та експлуатація 25

1.4 Підходи до виявлення шкідливого контенту 27

1.4.1 Ручний метод виявлення контенту 28

1.4.2 Автоматизований метод виявлення контенту 29

1.4.3 Порівняння ручного та автоматизованого методу виявлення контенту 30

2 ДОСЛІДЖЕННЯ ТА АНАЛІЗ МЕТОДІВ ТА ІНСТРУМЕНТІВ ДЛЯ МОНІТОРИНГУ ТА АНАЛІЗУ КОНТЕНТУ В TELEGRAM 31

2.1 Загальний огляд Telegram 31

2.1.1 Основні характеристики Telegram 31

2.1.2 Основні елементи взаємодії в Telegram 31

2.2 Регулювання шкідливого контенту в Telegram 33

2.3 Telegram API та його можливості для збору та аналізу даних 36

2.3.1 Використання Bot API для аналізу даних 36

2.3.2 Використання TDLib для аналізу даних 37

2.3.3 Використання Telethon для аналізу даних 37

2.3.4 Використання Pyrogram для аналізу даних 38

3 ДОСЛІДЖЕННЯ ТА АНАЛІЗ ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ ТА ТЕХНОЛОГІЇ OSINT ДЛЯ ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ 39

3.1 Застосування методів NLP для виявлення небезпечного контенту.. 39

3.2 Глибоке навчання у задачах розпізнавання шкідливих текстових повідомлень 40

3.2.1 Рекурентні та згорткові нейромережі (RNN, LSTM, CNN) 40

3.2.2 Трансформерні моделі (BERT, RoBERTa)..... 41

3.2.3 Моделі для багатомовних і кодозмішаних текстів 42

3.3 Моделі штучного інтелекту для виявлення шкідливого контенту у мультимедійних даних 43

3.4 OSINT як інструмент виявлення шкідливого контенту..... 45

3.4.1 Визначення поняття OSINT у контексті боротьби з шкідливим контентом 45

3.4.2 Загальні OSINT-інструменти для збору та обробки даних 46

3.4.3 Спеціалізовані OSINT-системи для виявлення шкідливого контенту..... 48

4 ПЕРСПЕКТИВИ ТА РЕКОМЕНДАЦІЇ ЩОДО РОЗВИТКУ ТЕХНОЛОГІЙ ДЛЯ МОНІТОРИНГУ TELEGRAM-КАНАЛІВ..... 50

4.1 Обґрунтування потреби в програмному рішенні 50

4.2 Обґрунтування використаних технологій та огляд розробленого рішення..... 51

4.3 Напрями подальшого розвитку 55

ВИСНОВКИ..... 58

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ 60

ДОДАТОК А 66

ДОДАТОК Б 67

ДОДАТОК В 71

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ ТА СИМВОЛІВ

AI, ШІ	– Artificial intelligence, Штучний інтелект
OSINT	– Open-source intelligence, Розвідка на основі відкритих джерел
IP	– Internet Protocol, Інтернет протокол
GPS	– Global Positioning System, Система глобального позиціювання
ООН	– Організація Об'єднаних Націй
API	– Application Programming Interface, Прикладний програмний інтерфейс
ЄС	– Європейський Союз
CSAM	– Child Sexual Abuse Materials, Матеріали із зображенням сексуального насильства над дітьми
ETIDAL	– Global Center to Combat Extremism, Глобальний центр боротьби з екстремістською ідеологією
TDLib	– Telegram Database Library, Бібліотека бази даних Telegram
SMS	– Short Message Service, Служба коротких повідомлень
HTTP	– Hypertext Transfer Protocol, Протокол передачі гіпертекстових документів
NLP	– Natural language processing, Обробка природної мови
LSTM	– Long short-term memory, Довга короткочасна пам’ять
CNN	– Convolutional Neural Networks, Згорткові нейронні мережі
BERT	– Bidirectional Encoder Representations from Transformers, Двоспрямовані кодувальні представлення з трансформерів
RNN	– Recurrent neural network, Рекурентна нейронна мережа
GRU	– Gated Recurrent Unit, Вентильний рекурентний вузол

C-BiLSTM – Convolutional Bidirectional LSTM, Згорткова двонаправлена LSTM

BLSTM – Bidirectional LSTM, Двонаправлена LSTM

mBERT – Multilingual BERT, Багатомовний BERT

XLM – Cross-Lingual Language Model, Міжлінгвістична мовна модель

XLM-R – XLM-RoBERTa, Міжлінгвістична мовна модель RoBERTa

ЗМІ – Засоби масової інформації

NLI – Natural Language Inference, Природна мовна інференція

HTML – HyperText Markup Language, Мова розмітки гіпертексту

URL – Uniform Resource Locator, Уніфікований локатор ресурсів

JSON – JavaScript Object Notation, Запис об'єктів JavaScript

ВСТУП

Світ швидко змінюється в результаті нових цифрових технологій, які революціонізували можливості та характер мереж і спілкування. Соціальні медіа – це соціальна взаємодія між людьми, під час якої вони створюють та обмінюються інформацією та ідеями у віртуальних спільнотах і мережах. Використання цифрових медіа та веб-платформ дозволяє користувачам активно обмінюватися інформацією, генерувати контент, співпрацювати та взаємодіяти один з одним [1]. Згідно з німецькою онлайн-платформою Statista, яка збирає та візуалізує дані, в 10 найпопулярніших соціальних мереж по всьому світу входять Facebook, YouTube, Instagram, WhatsApp, TikTok, WeChat, Facebook Messenger, Telegram, Snapchat, Douyin [2].

Загальноприйнято вважати, що платформи соціальних мереж представляють різноманітні думки, різні джерела новин, але, на жаль, й навмисно шкідливий контент [3].

Telegram створив у собі спільноту, яка закрита від суспільства, через це не підпадає під закон України “Про медіа”. Незважаючи на це, Telegram набирає все більше популярності серед українців через свою перевагу у швидкому донесенні інформації. Але в той же час, проблема полягає в тому, що цей месенджер не є прозорим з точки зору суспільства та держави, неможливо встановити власників каналів, фінансову звітність (багато каналів Telegram займаються збором коштів), в який спосіб відбувається співпраця із спільнотою, що створює ризики до маніпуляцій з громадськістю. Додатковою проблемою, але не менш важливою є те, що більшість російських активностей в умовах гібридної війни активно використовують анонімні Telegram-канали та поширює різні види шкідливого контенту серед населення України, через відсутність зворотного зв’язку з модерацією Telegram [4].

У зв’язку з цим виникає потреба у сучасних рішеннях, які можуть допомогти у боротьбі з шкідливим контентом. Зокрема, штучний інтелект (AI) – це технологія, яка дозволяє комп’ютерам і машинам імітувати людське навчання, розуміння, вирішення проблем, прийняття рішень, творчість і автономію [5]. У

свою чергу, OSINT (Open-Source Intelligence) – розвідка, отримана шляхом збору, оцінки та аналізу загальнодоступної інформації з метою відповіді на конкретне запитання розвідки [6]. Поєднання цих технологій надає нові можливості для автоматизації аналізу, моніторингу та ідентифікації шкідливого контенту, що потребує дослідження.

Метою кваліфікаційної роботи є аналіз існуючих інструментів для моніторингу та аналізу контенту в Telegram, а також дослідження використання штучного інтелекту та технологій OSINT для запобігання шкідливому контенту в інформаційному середовищі. На основі отриманих результатів – розробка інструменту для автоматичного виявлення небезпечного контенту у Telegram-каналах, який інтегрує розглянуті технології.

1 ОГЛЯД ТА АНАЛІЗ ТЕОРЕТИЧНИХ ОСНОВ ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ В ІНФОРМАЦІЙНОМУ СЕРЕДОВИЩІ

1.1 Визначення шкідливого контенту

Визначення шкідливого (або небезпечного) контенту можуть трактуватись різними способами, але сутність однакова. Всесвітньо відома організація Google інтерпретує небезпечний контент, як контент, який просуває шкідливі дії, заохочує до них або сприяє їх скоєнню [7]. Онлайн платформа кіберполіції України “КіберБрама” визначає, що шкідливий контент (небажаний контент) – це матеріали (зображення, відео, аудіо, тексти), що містять зображення насильства, порнографію, пропаганду наркотичних засобів, азартних ігор, також це різноманітні комп'ютерні віруси та шпигунські програми [8]. Громадська організація “STOP SEXТинг”, яка створена з метою безпеки українських дітей в Інтернеті, опублікувала, що шкідливий контент – це контент в Інтернеті, що приносить людині страждання чи шкоду, або може спонукати до небезпечних дій [9].

Як можемо побачити, шкідливий контент – це достатньо широкий термін, він охоплює багато різних речей. В цій роботі це визначення буде використовуватись в більш загальному вигляді: шкідливий контент – це будь-які матеріали в Інтернеті, які приносять шкоду. В залежності від контексту, синонімами до слова шкідливий контент будемо вважати слова: шкода, насильство, образа, зловживання та схожі.

1.2 Підхід до класифікації шкідливого контенту

Існує багато різних класифікацій шкідливого контенту. Розглянемо класифікацію, яка описана у статті [10]. При побудуванні типології шкідливого контенту автори дотримувались таких основних принципів:

- Уникати використання суб'єктивних прикметників як основних критеріїв оцінки. Це пояснюється тим, що такі недостатньо визначені або суб’єктивні фрази, як «ненависницький», «токсичний» або «такий, що змушує залишити розмову», без подальших пояснень, можуть бути інтерпретовані по-

різному в залежності від особи або системи, що здійснює класифікацію цих даних, які надалі називатимемо анотатором.

- Надавати перевагу більш вузько націленим категоріями замість тих, які охоплюють декілька типів поведінки. Прикладом, є поведінка, яку називають “токсичною” або “булінгом”. Вона може містити поєднання ненависті ідентичності, загальних образ або неприйнятної сексуальної лексики. Це спрощує анотацію класів за рахунок того, що дозволяє уникнути плутанину серед анотаторів та забезпечує пояснення, чому дані саме так класифіковані.
- Розгляд типу жертви, на яку спрямований шкідливий контент. Потрібно використовувати різни типи жертв та не змішувати їх в одному понятті, коли це можливо. Наприклад, замість того, щоб створювати загальну категорію для розпізнавання сексуально відвертого контенту, рекомендується окремо позначати контент, що має сексуальний підтекст і спрямований на конкретну людину, від контенту, який рекламує послуги для дорослих. Однак в деяких випадках, такий підхід важко застосувати, бо форма шкоди не може бути чітко розпізнана за типом жертви, прикладом цього є дезінформація, в такому випадку краще застосовувати тематичний підхід, який є зрозумілішим серед анотаторів та користувачів.
- Подумати про потенційні подальші дії. Якщо певний тип поведінки завжди має серйозні наслідки, наприклад, правові дії, не слід об’єднувати його з поведінкою, яка таких наслідків не має. Наприклад, контент, що стосується сексуального насильства над дітьми, є неприпустимим за жодних обставин і підлягає повідомленню правоохоронним органам, тоді як образи із сексуальними висловами, хоч і неприйнятні, навряд чи матимуть юридичні наслідки.

Незважаючи на те, що автори надають перевагу деталізації класів, їх не можна вважати взаємовиключними та й ієрархічне розташування типів не завжди можливе. Як наслідок, з’являються випадки, які можна віднести до декількох типів. Наприклад, “Час пристрелити цього н*****”, де останнє слово є расовим образою, слід класифікувати як «Атака на особистість» та «Загроза насильства». “Час розстріляти цю школу” – це погроза насильством без нападу на особистість. “Н****** тут не раді” – це ненасильницька атака на особистість.

Окрім основних принципів до класифікації автори зазначають, що важливо визначити тяжкість шкідливого контенту. Усі форми зловживань є проблематичними і потребують певних засобів для їх виявлення та усунення, щоб пом'якшити їхній вплив на користувачів. Незважаючи на те, що спричинена шкода може інтерпретуватися по-різному залежно від одержувача та контексту, деякі онлайн-насильства становлять безпосередню або довготривалу небезпеку для здоров’я людей або є порушенням закону. Тому для кожного типу образ слід навести кваліфікацію, яка визначає серйозність, інакше кажучи встановити умови, які визначають, що саме може вважатись серйозним насильством. Зрозуміло, що це потрібно для того, щоб екстремально швидко реагувати на критично серйозні зловживання. Автори пропонують чіткі критерії тяжкості до загрози:

- Використання мови, що виражає прямий намір (тяжка) проти використання пасивної мови або просто побажання (не тяжка).
- Загрози заподіяння шкоди, які чутливі до часу або є невідкладними загрозами вважаються тяжкими.
- Наслідки або ступінь шкоди, пов'язані з насильством, тобто дії, що призводять до смерті або довготривалої фізичної чи психологічної травми, вважаються тяжкими.
- Вразливість жертви, наприклад, напади, спрямовані на членів груп, які історично були маргіналізовані, дегуманізовані або об'єктивізовані, вважаються тяжкими.
- Порушення особистого життя та згоди вважаються тяжким.
- Порушення чинного законодавства, включно з міжнародно визнаними політиками, розглядаються як тяжкі.

1.3 Класифікація шкідливого контенту

Використовуючи методологію описану в пункті 1.2, автори статті [10] прийшли до такої класифікації, яка показана на рис. 1.1. Розглянемо детальніше кожен з типів шкідливого контенту.

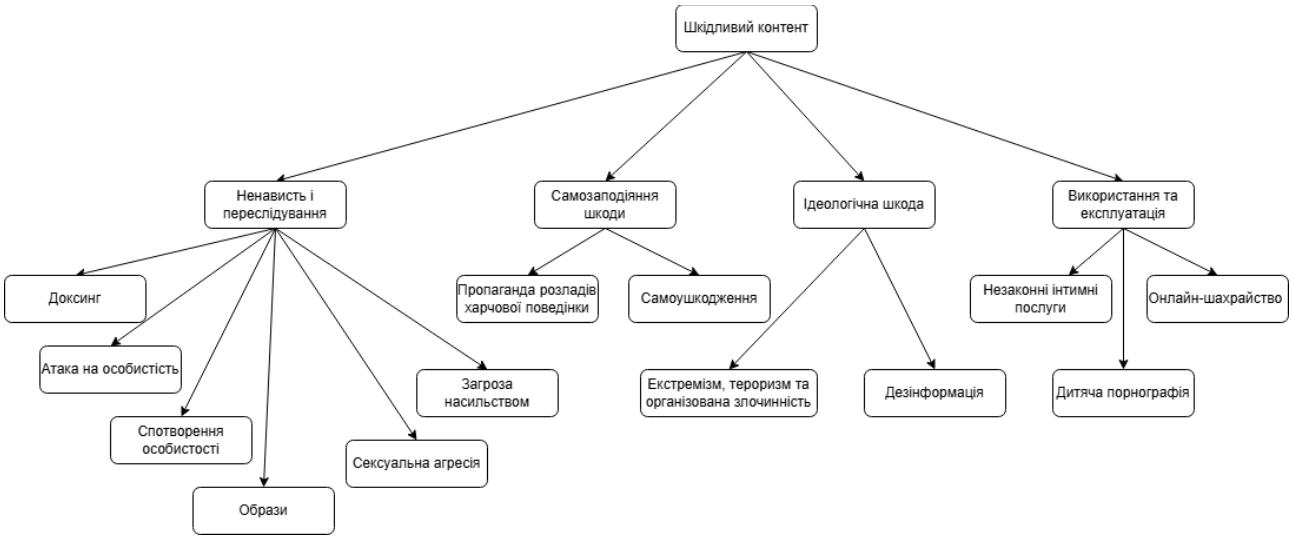

Рисунок 1.1 – Класифікація шкідливого контенту

1.3.1 Ненависть і переслідування

Ненависть і переслідування – це цілеспрямоване зловживання, спрямоване на конкретну особу або групу людей (наприклад, за ознакою ідентичності) з метою заподіяти їм шкоду: принизити, залякати, дискредитувати або зганьбити. Важливо зазначити, що висловлювання ненависті чи образи, спрямовані на організації, компанії, державні установи або абстрактні ідеї, такі як політичні ідеології, релігійні вчення чи соціальні концепції, не входять до цього визначення.

У цьому розділі розглянемо основні форми ненависті та переслідування: доксинг, атаки на особистість, спотворення особистості, образи, сексуальну агресію та загрози насильством.

Доксинг – це форма серйозного зловживання, коли зловмисник намагається завдати шкоди людині, публікуючи її особисту ідентифікаційну інформацію у відкритий доступ.

Під час атаки доксингу конфіденційні дані зазвичай поширюються на веб-сайтах, які дозволяють анонімне розміщення контенту та не вживають заходів для видалення матеріалів, що містять переслідування.

Доксинг може призвести до ще небезпечнішої форми атаки, відомої як сватинг – коли зловмисник робить неправдивий виклик правоохоронців, повідомляючи про насильство за адресою жертви. Це може спричинити

небезпечну ситуацію, наприклад, коли спеціальні підрозділи вибивають двері в будинок жертви та сприймають жертву за загрозу суспільству.

Особисті дані, які не можна розголошувати без згоди людини:

- Фізичне чи віртуальне місцезнаходження, наприклад, домашня або робоча адреса, IP-адреса чи GPS-координати.
- Контактна інформація, зокрема приватна електронна пошта та номери телефонів.
- Документи, що підтверджують особу, такі як номери соціального страхування, паспортні дані, державні чи студентські посвідчення.
- Цифрові ідентифікатори, включно з обліковими записами в соцмережах, нікнеймами у чатах і паролями.
- Фінансова інформація, наприклад, номери банківських рахунків або кредитних карток.
- Кримінальна та медична історія.

Проте публікація інформації, яка вже є у відкритому доступі, наприклад, місця навчання чи роботи, а також електронних адрес, які люди самі розміщують на своїх особистих сайтах, не вважається доксингом. Також це не стосується випадків, коли люди свідомо й добровільно діляться своїми особистими даними.

Атака на особистість – це форма онлайн-насильства, коли зловмисники здійснюють жорстокі напади на окремих людей або групи через їхню належність до захищених або вразливих спільнот.

Під час такої атаки кіберзлочинець використовує формулювання, що мають на меті дегуманізувати, переслідувати або сприяти насильству на основі ідентичності жертви. Це можуть бути образи, дискримінаційні висловлювання або навіть прямі погрози, хоча їхня присутність не є обов’язковою умовою.

Оскільки єдиного визначення мови ворожнечі немає ні в наукових дослідженнях, ні в правовому чи галузевому середовищі, поняття атак на особистість зазвичай формується навколо трьох основних принципів:

1) захист вразливих груп;

2) захист певних характеристик та ідентичності особи;

3) заборона мови ворожнечі без чіткого визначення її меж.

Багато платформ захищають користувачів від атак, спрямованих на їхню расу, національність, релігію, стать, вік, інвалідність, військовий статус або сексуальну орієнтацію. Деякі також додають захист для груп за соціально-економічним статусом, імміграційним статусом або станом здоров'я.

Автори [10] пропонують деталізований підхід, для визначення атак на особистість, розрізняючи серйозні атаки (наприклад, ті, що містять дегуманізацію або заклики до насильства) та менш агресивні прояви, такі як поширення негативних стереотипів або дезінформації про вразливі групи. Ще однією важливою відмінністю є те, що об'єктом атаки має бути конкретна людина або група людей, а не інституція чи організація. Наприклад, напад на людей, які сповідують певну релігію (наприклад, євреїв або мусульман), вважається атакою на особистість, тоді як критика самої релігії (наприклад, юдаїзму чи ісламу) не підпадає під це визначення. Таким чином, висловлювання на кшталт: "Ти заслуговуєш на евтаназію, брудний ****" або " ****и заслуговують на знищення" розглядаються як атака на особистість. Натомість фраза "**** – ця релігія має зникнути" не вважається атакою на особистість, оскільки спрямована на ідеологічну концепцію, а не на конкретних людей, які її сповідують.

Враховуючи вище сказане, наведемо узагальнені типи контенту, які підпадають під класифікацію атаки на особистість і вважаються тяжкими формами зловживання:

- Відверте використання образливих слів та принизливих прикметників, спрямованих проти певної ідентичної групи.
- Погрози насильством або заклики до заподіяння шкоди щодо конкретної групи.
- Заклики до виключення, домінування або обмеження прав, спрямовані на певну ідентичну групу.

- Дегуманізація групи – порівняння з тваринами, комахами, хворобами чи брудом, а також узагальнення щодо фізичної непривабливості, низького інтелекту, психічної нестабільності або моральної нікчемності.
- Висловлювання про перевагу однієї групи над іншою, що належить до вразливих або захищених категорій.
- Відкрите зізнання у ненависті та нетерпимості до членів певної ідентичної групи.
- Заперечення чи невизнання ідентичності іншої людини, заклики до репаративної (конверсійної) терапії, використання імені, даного при народженні трансгендерній особі.
- Підтримка груп ненависті, які пропагують або поширюють вищезазначені форми агресії.

Наступні випадки не вважаються атаками на особистість:

- Критика установ або організацій, якщо вона не спрямована безпосередньо на людей, які до них належать.
- Поширення стереотипів, страхів чи дезінформації про певні групи (це розглядається окремо як спотворення особистості).
- Вживання образливих слів у середині спільноти, коли люди з групи використовують їх між собою, переосмислюючи їхнє значення.
- Освітні або мовознавчі дискусії про лайку та мову ворожнечі.
- Цитування чи аналіз висловлювань інших людей, якщо вони не є безпосередніми учасниками розмови.

Викривлення ідентичності – це висловлювання або твердження, які спотворюють дійсність, поширюють негативні стереотипи та узагальнення про захищені або вразливі групи.

Як і у випадку з атаками на особистість, до захищених груп належать люди, визначені за такими ознаками: вік, інвалідність, етнічна приналежність, гендерна ідентичність, військовий статус, національність, раса, релігія, сексуальна орієнтація. Додатковий захист можуть отримати іммігранти, люди з низьким соціально-економічним статусом або ті, хто має медичні захворювання.

Однак викривлення ідентичності не є настільки серйозним, як атака на особистість. Такі висловлювання часто подаються як факти, хоча можуть не мати доказової бази або бути прихованими формами суб'єктивної думки.

Ознаками спотворення особистості є:

- Поширення негативних стереотипів та узагальнень про захищені або вразливі групи, які не містять прямої дегуманізації чи тверджень про меншовартість.
- Висвітлення стереотипів як абсолютної правди, без доказів чи обґрунтування.
- Приховані форми дискримінації, тобто тонкі прояви упередженості щодо певної групи.
- Намір викликати страх або негативне ставлення до певної соціальної групи, без прямих закликів до насильства.

Однак висловлювання, що містять пряму образу або належать до категорії атаки на особистість, не розглядаються як спотворення особистості.

Образа – це висловлювання, спрямоване безпосередньо на людину або групу людей з метою принизити, спровокувати або висловити негативне ставлення. Вона може містити грубі слова, лайку або принизливі зауваження, але не стосується ідентичності людини, тобто її раси, національності, релігії, гендеру тощо. Якщо у висловлюванні використовуються дискримінаційні вислови або образливі характеристики, пов'язані з ідентичністю людини, то це вже атака на особистість, а не просто образа.

Образою вважаються такі дії, за умови, що жертва бере участь в розмові:

- Прямі образи, лайка або принизливі висловлювання, що не стосуються ідентичності людини (наприклад, інтелекту, характеру, емоційного стану).
- Знущання з особистих якостей, поглядів чи емоційного стану людини.
- Приниження зовнішності, бодішеймінг, засудження сексуального чи романтичного минулого.
- Знущання над людьми через їхній досвід насильства або знущань у минулому.

- Заклики до інших приєднатися до образливих висловлювань щодо певної особи.
- Редагування зображень з метою принизити людину.

Наступне не слід вважати прикладами образ:

- Образи, що ґрунтуються виключно на приналежності людини до захищеної групи, включаючи використання образливих слів – такі випадки розглядаються як атака на особистість, а не просто образа.
- Образи, спрямовані на людей, які не беруть участі в розмові, наприклад, знаменитостей.
- Самоіронія або самопринижения, коли людина жартує над собою.
- Образи, спрямовані на неживі предмети (наприклад, лайка на адресу гаджетів, техніки чи об’єктів).
- Освітній або дослідницький контент про домагання та мову ворожнечі, який не має на меті образити когось особисто.

Сексуальна агресія – це форма особистого зловживання, яка включає небажані сексуальні домагання, неприйнятну сексуалізацію, несхвалене поширення сексуального контенту та інші небажані сексуальні контакти.

До цього виду шкідливого контенту належать:

- Погрози або опис сексуальних дій, фантазій чи актів примусового сексу щодо конкретної особи.
- Небажані відверті описи сексуального характеру, навіть якщо вони стосуються самої людини, яка їх висловлює.
- Сексуальна об'єктивація, небажані домагання чи коментарі, які мають на меті принизити людину з сексуальної точки зору.
- Заклики або пропозиції некомерційних сексуальних взаємодій.
- Нав’язливі запити про оголені або сексуально відверті фото чи відео.
- Несанкціоноване поширення контенту, що зображує людину оголеною або під час сексуальних дій (включаючи фейкові зображення, наприклад, "порнопомста").

- Публікація фото чи відео, що демонструють інтимні частини тіла людини, навіть якщо вона одягнена або перебуває у громадському місці, без її згоди.
- Погрози розголошення інтимних зображень, переписок чи іншої конфіденційної інформації сексуального характеру.

До сексуальної агресії не відноситься:

- Обговорення визначень сексуальних термінів.
- Дискусії щодо сексуального здоров'я та добробуту.
- Використання сексуальних термінів без графічного чи принизливого підтексту.
- Образи, що містять сексуальні вирази, але не мають загроз чи домагань.
- Флірт, компліменти або романтичні коментарі, які не є сексуально відвертими або принизливими.

Загроза насильством. Багато онлайн-платформ забороняють користувачам висловлювати бажання завдати фізичної шкоди чи вбити іншу людину. Також не допускаються висловлювання, що прославляють, заохочують або виправдовують насильницькі дії, оскільки вони можуть спровокувати інших на вчинення насильства.

Погроза насильством охоплює контент, що містить хоча б один із таких елементів:

- Бажання фізично зашкодити людині чи групі людей, включно з погрозами сексуального насильства.
- Заклики до вбивства, нанесення серйозних травм або спричинення важкої хвороби.
- Підбурювання інших до самогубства чи самопошкодження.
- Заклики до насильницьких дій.
- Прославлення насильства або насильницьких подій.

Не є прикладами цього класу:

- Опис особистого досвіду насильства без його прославлення.
- Історичні описи чи дослідження насильства.

- Гіперболічні або метафоричні вислови, які не передбачають реальної загрози.

1.3.2 Самозаподіяння шкоди

Самозаподіяння шкоди включає в себе різні форми небезпечної поведінки, як фізичної, так і психологічної, спрямованої проти себе. Виявлення такого контенту необхідне для того, щоб ідентифікувати людей у стані кризи, надати їм підтримку та запобігти поширенню небезпечних практик у онлайн-спільнотах. В цьому контексті розглядаються дві основні форми самозаподіянної шкоди: пропаганда розладів харчової поведінки, самоушкодження.

Пропаганда розладів харчової поведінки. Розлади харчової поведінки – це психічні захворювання, які супроводжуються аномальними харчовими звичками та спотвореним ставленням до їжі. Багато онлайн-платформ забороняють контент, що заохочує або підтримує подібні розлади, щоб запобігти поширенню небезпечної поведінки. Автори [10] наводять такі типи небезпечного контенту, що стосується розладів харчової поведінки:

- Популяризація розладів харчової поведінки як прийнятного або бажаного способу життя (наприклад, контент, що пропагандує анорексію або булімію).
- Ідеалізація крайньої худорлявості та виснаженого вигляду тіла.
- Дискредитація висококалорійної їжі та висміювання людей із надмірною вагою з метою викликати відразу.
- Розповсюдження шкідливих порад та методів екстремального схуднення, які можуть завдати серйозної шкоди здоров’ю.

Такі речі, як наукові дослідження, освітні матеріали та просвітницькі кампанії, що стосуються розладів харчової поведінки, не є пропагандою розладів харчової поведінки. Окрім них, також не відносяться до цієї категорії: обговорення шляхів відновлення та ресурсів для допомоги людям із розладами харчової поведінки, особисті історії людей, які пережили розлад, якщо вони не прославляють цей стан.

Самоушкодження – це навмисне завдання собі фізичної шкоди за допомогою різних методів, таких як порізи гострими предметами, опіки, укуси, виривання волосся тощо. Люди, які вдаються до такої поведінки, роблять це як спосіб справлятися з емоційним болем чи стресом.

До самопошкоджувального контенту належать:

- Обговорення недавніх чи поточних випадків навмисного заподіяння шкоди собі.
- Висловлювання про суїцидальні думки, опис плану самогубства або заяви про намір накласти на себе руки.
- Прохання про інструкції щодо самопошкодження або способів приховати його сліди.
- Опис емоцій чи симптомів психічних розладів, що безпосередньо пов’язані із самоушкодженням, або обговорення тригерних подій, що призвели до такої поведінки.
- Просування чи підтримка поведінки, пов’язаної з самопошкодженням.

До самоушкодження не відноситься:

- Розповіді про особистий досвід одужання та лікування, які не пропагують самопошкодження.
- Обговорення способів подолання суїцидальних думок та прагнення до самопошкодження.
- Надання підтримки людям, які стикаються з думками про самопошкодження або вже мали подібний досвід.
- Наукові дослідження та освітні матеріали, спрямовані на запобігання самопошкодженню та суїциду.
- Обговорення депресії та психічних розладів, якщо вони не пов’язані безпосередньо з темою самопошкодження чи самогубства.

Розпізнавання контенту про самопошкодження важливе для надання підтримки людям, які перебувають у кризі, а також для запобігання поширенню потенційно небезпечних практик у спільнотах.

1.3.3 Ідеологічна шкода

Ідеологічна шкода – це поширення переконань, які можуть призвести до негативних наслідків для суспільства у довгостроковій перспективі. Контент, що належить до цієї категорії, може не мати чітко визначеної людської цілі на момент створення, проте він може сприяти суспільним кризам або порушенням безпеки. Наприклад, це можуть бути висловлювання, що відкрито ставлять під сумнів державну політику чи заходи охорони здоров’я, що може спричинити суспільні потрясіння, або схвалення ідеологій, пов’язаних із злочинністю, насильством чи дискримінацією. Нижче наведено визначення двох основних форм ідеологічної шкоди: екстремізму, тероризму та організованої злочинності, а також дезінформації.

Екстремізм, тероризм, організована злочинність. Визначення терміну «тероризм» – питання проблемне, оскільки в наш час існує понад 100 визначень цього явища [11]. Генеральна Асамблея ООН визначає його як “кримінальні дії, спрямовані на створення стану терору серед населення, групи осіб або окремих людей з політичних мотивів, які за жодних обставин не можуть бути виправдані, незалежно від політичних, філософських, ідеологічних, расових, етнічних, релігійних чи інших аргументів, що використовуються для їх виправдання” [12]. Законодавство України визначає, що “тероризм – суспільно небезпечна діяльність, яка полягає у свідомому, цілеспрямованому застосуванні насильства шляхом захоплення заручників, підпалів, убивств, тортур, залякування населення та органів влади або вчинення інших посягань на життя чи здоров'я ні в чому не винних людей або погрози вчинення злочинних дій з метою досягнення злочинних цілей” [13]. Тоді як терористичні групи переважно асоціюються з насильницькою діяльністю, екстремізм може включати як насильницькі, так і ненасильницькі форми вираження. Організовані злочинні угруповання, які часто вдаються до насильства, зазвичай не мають політичних чи ідеологічних цілей, а керуються економічними інтересами [10].

Шкідливий контент, пов'язаний з екстремізмом, тероризмом та організованою злочинністю:

- Вербування до терористичної організації, екстремістського угруповання чи організованої злочинної групи.
- Вихваляння та популяризація організованої злочинності, терористичних чи екстремістських угруповань або дій, вчинених такими угрупованнями.
- Сприяння терористичній організації, екстремістському угрупованню чи організованій злочинній групі.
- Контент, що містить символіку, яка, як відомо, представляє терористичну організацію, екстремістське угруповання або організовану злочинну групу.

Дезінформація – це хибна або оманлива інформація. Вона може поширюватися користувачами, які не усвідомлюють її неправдивості та не мають на меті завдати шкоди.

Дезінформація є підвидом фейкової інформації, але відрізняється тим, що її навмисно поширюють з метою маніпуляції. Намір дезінформації зазвичай є зловмисним – наприклад, для підриву довіри до людини чи організації, отримання політичної або фінансової вигоди.

До цієї категорії належать:

- Фейкові новини – вигадані події, що подаються як справжні.
- Помилкові чутки – неправдиві відомості, що швидко поширюються.
- Теорії змови – ідеї, які безпідставно стверджують існування таємних змов.
- Фальшиві сенсації та розіграші.
- Маніпулятивні коментарі та відгуки, зокрема підроблені відгуки про продукти.

Багато онлайн-платформ все частіше забороняють певні форми дезінформації, серед яких:

- Медично необґрунтовані заяви, що загрожують здоров’ю та безпеці, наприклад, фейкові ліки, неправдива інформація про пандемії чи надзвичайні ситуації у сфері охорони здоров'я.

- Хибна або маніпулятивна інформація про вразливі або захищені групи людей.
- Неправдива інформація про вибори, що може порушити чесність виборчого процесу або завадити громадській участі.
- Теорії змови, які спотворюють реальність або створюють паніку.
- Заперечення добре задокументованих історичних подій.
- Фальшиві відгуки та коментарі, що вводять людей в оману.
- Навмисне викривлення фактів для підриву довіри або завдання шкоди.
- Маніпуляція аудіо- та візуальним контентом.

1.3.4 Використання та експлуатація

Щоб забезпечити безпечний онлайн-простір, у цифрових спільнотах не допускається контент, створений з метою отримання вигоди шляхом завдання шкоди іншим – фінансової, сексуальної чи фізичної. До форм експлуатації належить незаконні інтимні послуги, дитяча порнографія та шахрайство.

Форми незаконних інтимних послуг переходять межу незаконної діяльності та часто експлуатують вразливих людей, тому вони класифікуються як тяжкі форми зловживання.

До інтимних послуг належать:

- Реклама або пропозиція незаконних інтимних послуг, зокрема проституції, ескорт-послуг, оплачених фетиш- чи домінування-сесій, еротичних масажів.
- Організація торгівлі людьми для сексуальної експлуатації.
- Вербування для участі у секс-шоу, сексуальних чатах або інших формах онлайн-експлуатації.

Дитяча порнографія охоплює будь-який контент, що зображує сексуальне насильство та експлуатацію осіб, які не досягли 18 років. Це включає не лише зображення та відео, а й текстовий контент, що має експлуатаційний характер.

Дитяча порнографія є тяжкою формою зловживання та включає:

- Зображення та відео, які зображують неповнолітніх у порнографічному, сексуально провокаційному чи насильницькому контексті, включаючи ілюстровані або цифрово змінені матеріали.
- Розповсюдження матеріалів для дорослих або матеріалів дитячої порнографії серед неповнолітніх.
- Встановлення довірливих відносин з дитиною з метою подальшої сексуальної експлуатації.
- Сексуально неприйнятні висловлювання, адресовані неповнолітнім.
- Організацію реальних сексуальних контактів або прямі запити на сексуальні матеріали від неповнолітніх.
- Поширення порад або виправдання сексуального насильства над дітьми.

Онлайн-шахрайство – це спроби обманним шляхом змусити людину передати гроші або конфіденційну інформацію, використовуючи маніпулятивні або нав'язливі тактики. Шахраї можуть поступово вибудовувати фальшиві відносини з жертвою або видавати себе за авторитетну особу чи експерта.

Види шахрайства, заборонені у цифрових спільнотах:

- Обманні схеми, що змушують користувачів передавати гроші або розкривати особисту інформацію (наприклад, фішинг).
- Обіцянки отримання великих сум грошей в обмін на початковий внесок через банківський переказ, подарункові картки або передплачені дебетові картки.
- Оголошення про виграш у лотереї або обіцянки грошових винагород і подарунків (лотерейні шахрайства).
- Шахрайство з підробленими романтичними або військовими профілями.
- Фальшиві пропозиції про списання боргів або покращення кредитної історії.
- Вербування в пірамідальні схеми.

1.4 Підходи до виявлення шкідливого контенту

З огляду на величезний обсяг контенту, створеного користувачами на різних платформах соціальних медіа, виявлення та модерація шкідливого контенту набуває критично важливого значення [14]. Коли контент публікується на платформі соціальних медіа, він проходить перевірку, щоб визначити, чи є він шкідливим чи нешкідливим. Рис. 1.2 показує основні етапи виявлення та модерації контенту в соціальних мережах.

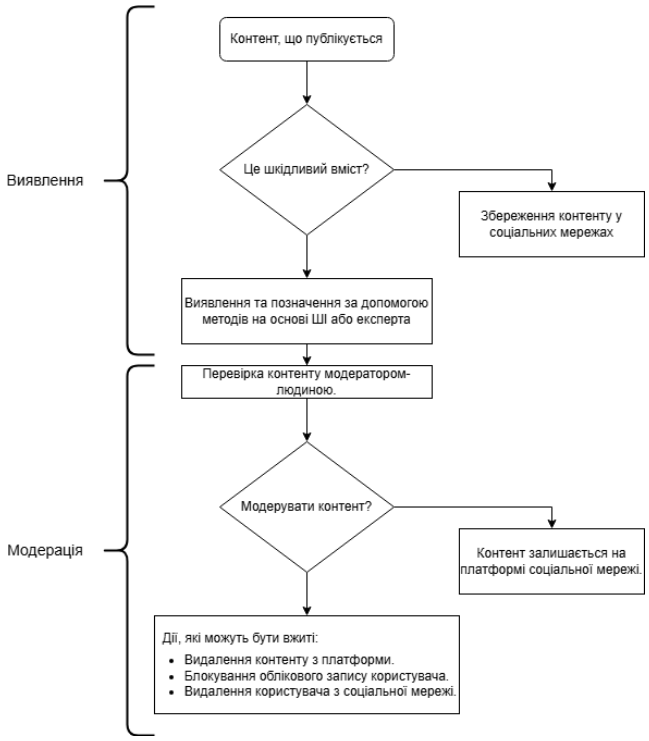

Рисунок 1.2 – Виявлення та модерація контенту в соціальних мережах

Виявлення шкідливого контенту – це завдання з ідентифікації неприйнятного або потенційно небезпечного матеріалу, розміщеного користувачами на платформах соціальних мереж. Під модерацією розуміється процес оцінки та перевірки виявленого контенту на відповідність встановленим правилам і політикам конкретної платформи соціальних медіа [14].

Перед тим як модератори соціальних мереж можуть прийняти рішення щодо видалення або блокування контенту, його спочатку потрібно ідентифікувати та класифікувати як шкідливий або нешкідливий. В основному існує два способи виявлення контенту це ручний спосіб (людська модерація) та автоматизований (ШІ-модерація). Зважаючи на величезну кількість контенту, що

щоденно публікується в соціальних мережах, ручні методи перевірки є нерентабельними та неефективними. Штучний інтелект став ключовим інструментом для автоматичного ідентифікації та фільтрації шкідливого контенту. Методи ШІ, що застосовуються для виявлення шкідливого контенту, включають машинне навчання, глибоке навчання та обробку природної мови [14].

Як вже було згадано раніше, процес модерації контенту в онлайн-середовищі може здійснюватися двома основними методами: ручним та автоматизованим. Обидва методи мають свої переваги та обмеження, і їх застосування залежить від конкретних потреб платформи, кількості контенту та рівня складності його оцінки. Розглянемо їх детальніше

1.4.1 Ручний метод виявлення контенту

Ручні методи виявлення та аналізу контенту є системний підходом, який передбачає перевірку даних вручну людиною без використання автоматизованих інструментів [15].

Ручна модерація контенту передбачає, що людський модератор безпосередньо переглядає повідомлення, оцінює їхню відповідність до політик платформи та ухвалює рішення про їхнє збереження або видалення. Людські модератори можуть не лише блокувати неприйнятний контент, але й сприяти розвитку здорової дискусії, коригуючи поведінку користувачів та пояснюючи правила платформи [16].

Процес ручної модерації зазвичай включає кілька ключових етапів. Спочатку модератор переглядає контент та оцінює його відповідність правилам спільноти. Далі він визначає, чи порушує ця публікація закони та політику компанії. Окрім безпосереднього аналізу тексту, модератор також бере до уваги контекст повідомлення та може оперативно реагувати на сарказм, іронію, культурні особливості та приховані форми агресії, що є важливою перевагою перед автоматизованими системами [16].

Однак ручний метод має і свої недоліки. По-перше, він є повільним і трудомістким, оскільки модератори можуть обробляти лише обмежену кількість

повідомлень за певний час. По-друге, психологічний тиск на модераторів є дуже високим, адже вони щодня стикаються зі шокуючим або травматичним контентом. Це може призводити до емоційного вигорання та зниження об’єктивності ухвалених рішень. По-третє, цей метод є дорогим для компаній, оскільки потребує постійного залучення, навчання та підтримки великої кількості модераторів [16].

1.4.2 Автоматизований метод виявлення контенту

Автоматизована модерація використовує алгоритми штучного інтелекту для аналізу контенту та ухвалення рішень щодо його прийнятності. Такі системи в основному працюють на основі машинного навчання, тобто тренуються на основі попередніх модераційних рішень, ухвалених людьми [16].

Процес автоматизованої модерації включає кілька етапів. Спочатку система збирає великі обсяги даних та аналізує закономірності в модераційних рішеннях, ухвалених людьми. Далі вона оцінює текст або зображення за різними критеріями, такими як використані слова, контекст або історія подібних повідомлень. Якщо контент перевищує встановлений поріг «небажаності», він може бути автоматично заблокований або позначений для подальшої перевірки людиною [16].

Основною перевагою автоматизованої модерації є її швидкість та масштабованість. ШІ може аналізувати тисячі повідомлень щосекунди, що дозволяє платформам швидко реагувати на порушення та підтримувати чистоту дискусійного простору. Окрім цього, автоматизовані системи зменшують навантаження на людських модераторів, допомагаючи їм уникнути постійного контакту з травматичним контентом [16].

Однак у цього підходу є й суттєві недоліки. Машина має свої обмеження в інтерпретації контенту; оскільки їм бракує емпатії, почуття гумору, розуміння іронії чи сарказму. Вони лише повторюють рішення, ухвалені модераторами-людьми, покірно виконуючи завдання з видалення контенту на основі раніше отриманих навчальних даних. Коли змінюються мовні конструкції та з’являються нові форми неприйнятних виразів, людські модератори швидко адаптуються,

розпізнають контекст та значення повідомлень. Натомість ШІ пристосовується повільніше, оскільки він поступово оновлює свої моделі. Це означає, що машина може бути застарілою та не відповідати сучасним стандартам модерації контенту [16].

1.4.3 Порівняння ручного та автоматизованого методу виявлення контенту

У табл. 1.1 показані переваги та недоліки двох розглянутих методів виявлення та аналізу контенту.

Таблиця 1.1 –Порівняння штучного інтелекту та людини у виявленні та аналізі контенту

Категорія	Метод виявлення	
	Ручний (людина)	Автоматичний (штучний інтелект)
Переваги	• Швидко адаптується • Має емпатію • Чутливий до контексту	• Працює у реальному часі (24/7) • Ефективний при обробці великих обсягів даних • Стабільний у дотриманні правил • Не страждає від психологічного навантаження
Недоліки	• Має обмеження за швидкістю та кількістю оброблених повідомлень • Суперечливий у прийнятті рішень • Вразливий до жорстокого контенту	• Повторює раніше ухвалені рішення та повільно адаптується • Погано розпізнає гумор, іронію, сарказм

Таким чином згідно з таблицею, можна зробити висновок про доцільність використання гібридного підходу, який поєднує переваги обох методів. Гібридний підхід дозволяє максимально використовувати сильні сторони кожного з методів. Штучний інтелект ефективно обробляє великі обсяги даних та працює безперервно, але має обмеження у розпізнаванні гумору та контексту. Людина, в свою чергу, краще справляється з оцінкою емоційного контексту та адаптацією до нових ситуацій.

2 ДОСЛІДЖЕННЯ ТА АНАЛІЗ МЕТОДІВ ТА ІНСТРУМЕНТІВ ДЛЯ МОНІТОРИНГУ ТА АНАЛІЗУ КОНТЕНТУ В TELEGRAM

2.1 Загальний огляд Telegram

2.1.1 Основні характеристики Telegram

Telegram – це швидка та безпечна програма для обміну повідомленнями, яка підтримує одночасну роботу на різних пристроях. У Telegram можна надсилати текстові повідомлення, зображення, відео та файли різних форматів (наприклад, doc, zip, mp3), а також створювати групи з кількістю учасників до 200 тисяч осіб або канали для трансляції контенту без обмежень щодо кількості підписників. Завдяки відкритому API, користувачі можуть створювати боти для автоматизації різних процесів [17].

Крім того, Telegram забезпечує наскрізне шифрування для голосових і відеодзвінків, а також підтримує голосові чати, в яких можуть брати участь тисячі користувачів. Протокол MTProto, розроблений спеціально для Telegram, поєднує в собі високу швидкість передачі даних та надійність, що робить обмін повідомленнями захищеним навіть на слабких з’єднаннях [17].

Для підвищення безпеки платформи Telegram пропонує користувачам два основні рівні шифрування:

1) Шифрування між сервером і клієнтом в хмарних чатах (особистих та групових).

2) Наскрізне шифрування в секретних чатах, яке гарантує, що повідомлення не зберігаються на серверах Telegram.

2.1.2 Основні елементи взаємодії в Telegram

Telegram пропонує різноманітні формати взаємодії користувачів, які можна розділити на приватні (або секретні) чати, групи, канали, боти.

Приватні чати дозволяють користувачам обмінюватися текстовими повідомленнями, медіа-файлами та голосовими чи відеоповідомленнями. Повідомлення зберігаються в хмарі Telegram і можуть бути доступні з різних пристроїв. Користувачі можуть видаляти повідомлення як для себе, так і для

співрозмовника. Секретні чати є покращеними стосовно безпеки приватними чатами. Секретні чати використовують наскрізне шифрування, що забезпечує високу конфіденційність обміну повідомленнями. Ці чати не зберігаються на серверах Telegram, а їхні повідомлення зникають після прочитання або за певним таймером. Також повідомлення з секретних чатів не можна пересилати іншим користувачам [17].

Групи в Telegram – це чати, які можуть об’єднувати до 200 000 учасників, забезпечуючи зручну взаємодію між користувачами. Вони дозволяють обмінюватися повідомленнями, файлами, реагувати на публікації та використовувати згадки й відповіді для ефективного спілкування. Користувачі можуть редагувати або видаляти повідомлення, що автоматично зникають у всіх учасників. Групи доступні на будь-якому пристрої, а завдяки миттєвому пошуку можна швидко знаходити потрібну інформацію. Для зручної навігації Telegram підтримує згадки, відповіді та хештеги, що спрощує участь у великих дискусіях. Якщо група надто активна, можна налаштувати сповіщення, отримуючи лише важливі повідомлення. Адміністратори можуть закріплювати повідомлення, щоб вони залишалися у верхній частині чату [17].

Канали – це інструмент для трансляції повідомлень великій аудиторії. Вони дозволяють власникам поширювати інформацію без обмежень щодо кількості підписників. На відміну від груп, де учасники можуть обмінюватися повідомленнями, у каналах публікувати контент можуть лише адміністратори. Коли адміністратор створює новий допис у каналі, він автоматично підписується іменем каналу, а не окремого користувача. Це забезпечує анонімність авторів та робить канали ефективним засобом для офіційних оголошень, новин та масового інформування. Нові підписники мають доступ до всієї історії повідомлень, що дозволяє їм ознайомитися з попереднім контентом одразу після приєднання [18].

Боти – це спеціальні програми, які працюють безпосередньо у Telegram і виконують різноманітні автоматизовані завдання. Вони створюються сторонніми розробниками за допомогою Telegram Bot API і можуть використовуватися для різних цілей: від розважального контенту до модерації груп, пошуку інформації

або автоматизації рутинних дій. Боти можуть працювати як у приватних чатах, так і у групах, реагуючи на команди користувачів або виконуючи завдання у фоновому режимі. Вони не мають доступу до особистих даних користувачів, окрім інформації, яку їм безпосередньо надсилають. Для створення бота потрібні базові навички програмування, проте в Telegram доступні вже готові рішення, які можна використовувати без додаткових налаштувань [17].

2.2 Регулювання шкідливого контенту в Telegram

Оскільки Telegram є відкритою та безкоштовною платформою для обміну повідомлення, очевидно, що в ньому також є великий ризик розповсюдження шкідливого контенту. Регулювання такого контенту в цьому мессенджері відбувається на основі умов обслуговування, які пропонують Telegram. В цих умовах вказано, що користувач при реєстрації облікового запису погоджується не:

- Використовувати сервіс для надсилання спаму та шахрайства;
- Пропагувати насильство;
- Розміщувати незаконний порнографічний контент;
- Займатися діяльністю, яка визнана незаконною в більшості країн. Це включає жорстоке поводження з дітьми, продаж або пропонування незаконних товарів і послуг (наркотики, вогнепальна зброя, підроблені документи), розголошення особистих даних інших з метою залякування або знущання над ними (доксинг) тощо.

В основному ці правила встановленні відповідно до Регламенту (ЄС) 2022/2065 Європейського Парламенту та Ради від 19 жовтня 2022 року про єдиний ринок цифрових послуг (Digital Market Act) і внесення змін до Директиви 2000/31/EC (Digital Services Act) [19, 20].

Відповідно до того ж регламенту, для запобігання зловживанню своєю відкритою платформою Telegram використовує поєднання модерації загальнодоступного контенту на основі штучного інтелекту та ручної перевірки, а також обробляє скарги користувачів. Платформа надає пріоритет мінімально необхідним обмежувальним заходам, прагнучи створити безпечне цифрове

середовище та ефективно протидіяти шкідливому контенту. Хоча певні обмеження можуть застосовуватися або зніматися автоматично, критично важливі рішення ухвалюються лише після перевірки модератором [20].

Telegram може тимчасово або назавжди обмежити певні функції облікового запису, наприклад, заборонити взаємодію з користувачами, які, ймовірно, не знайомі з власником облікового запису, або обмежити можливість створення та участі в публічних спільнотах. Облікові записи, боти та групи, що видають себе за інших осіб або намагаються вводити користувачів в оману, можуть отримати позначку FAKE або SCAM, яка відображатиметься в їхньому загальнодоступному профілі. У разі серйозних порушень модератори можуть заблокувати або повністю видалити облікові записи, боти, публікації, канали та групи.

Якщо законодавство дозволяє, Telegram повідомляє користувачів про застосовані обмеження та надає інструкції щодо їх можливого скасування. Деякі обмеження можуть бути зняті автоматично. У разі помилкового блокування Telegram перегляне рішення та оновить або скасує обмеження, надіславши відповідне сповіщення [20].

Як вже було згадано, Telegram використовує ручну модерацію в поєднанні з автоматизованим виявленням завдяки штучному інтелекту і за допомогою скарг звичайних користувачів. Окрім проактивних заходів та скарг користувачів, Telegram обробляє тисячі повідомлень про CSAM-контент від глобальних сторонніх організацій через автоматизовані адреси блокування:

- abuse@telegram.org;
- stopCA@telegram.org.

Ручна модерація передбачає розгляд скарг користувачів, повідомлень від організацій, виявлених порушень модераторами Telegram або штучним інтелектом та ухваленні відповідних рішень щодо контенту або облікових записів [20].

Telegram надає статистику роботи модерації. Наприклад, на рис. А.1 в додатку А показано загальну кількість заблокованих груп і каналів у період з 13 лютого до 15 березня 2025 року [21].

Крім того, у партнерстві з такими організаціями, як ETIDAL та Глобальний центр протидії екстремістській ідеології, Telegram бореться з терористичним контентом. Зокрема, на рис. А.2 в додатку А представлено кількість заблокованих терористичних спільнот у той самий період [21].

Telegram дотримується політики нульової толерантності до матеріалів, що містять зображення сексуального насильства над дітьми. У табл. 2.1 наведено список чотирьох найбільших некомерційних організацій, які у 2024 році подали звіти про такий контент для його блокування в Telegram [21].

Таблиця 2.1 – Кількість заблокованих матеріалів сексуального насильства над дітьми

Організація	Кількість заблокованих матеріалів сексуального насильства над дітьми	
	Січень – Червень 2024	Липень – Листопад 2024
Stichting Offlimits	2142	21765
Canadian Centre for Child Protection	7131	5902
National Center for Missing & Exploited Children	1933	3082
Internet Watch Foundation	107	141
Загальна кількість	11313	30890

Окрім загальної модерації, у групових чатах працюють адміністратори, яких призначає власник групи або каналу. Вони мають можливість встановлювати власні правила чату, аналізувати повідомлення та стежити за їх дотриманням. Для автоматизації виявлення та редагування шкідливого контенту адміністратори використовують ботів. Функціонально боти нагадують звичайних користувачів, проте виконують запрограмовані модераційні завдання [17].

2.3 Telegram API та його можливості для збору та аналізу даних

API або ж інтерфейс програмування застосунків є набором визначених методів для швидкої розробки програмного забезпечення. Telegram надає три типи API для розробників:

- Bot API дозволяє легко створювати програми, які використовують повідомлення Telegram як інтерфейс.
- TDLib дозволяє створювати власні клієнти Telegram.
- Gateway API дозволяє будь-якій компанії, програмі чи веб-сайту надсилати коди підтвердження через Telegram замість традиційних SMS [22].

Серед офіційних API Telegram для збору та аналізу даних найпоширенішими є Bot API і TDLib. Перший підходить для створення ботів з обмеженим доступом до контенту, другий – для повноцінної роботи з історією повідомлень.

Окрім офіційних, існують зручні сторонні бібліотеки, такі як Telethon і Pyrogram, що активно використовуються для автоматизації, аналізу даних Telegram.

Нижче розглянемо особливості кожного з них.

Для аналізу потенційно шкідливого контенту у Telegram можна інтегрувати сторонні інструменти, зокрема алгоритми штучного інтелекту та методи обробки природної мови. Вони дозволяють автоматично виявляти заборонені теми, мову ворожнечі, дезінформацію, спам або контент, що порушує правила спільнот.

2.3.1 Використання Bot API для аналізу даних

Telegram Bot API – це HTTP-інтерфейс, який дозволяє створювати ботів для автоматизації взаємодії з користувачами в Telegram [23]. Однією з ключових його функцій є можливість отримання повідомлень та їхнього змісту. Бот може отримувати текстові та мультимедійні повідомлення від користувачів у приватних чатах, групах або каналах [23].

Важливою функцією API є фільтрація повідомлень. Бота можна налаштовувати на визначення ключових або заборонених слів, що дозволяє блокувати неприйнятний контент і стежити за дотриманням правил [23].

Попри широкі можливості Telegram Bot API, його використання для аналізу даних має низку обмежень. Бот отримує повідомлення лише за умови, що його додано до групи або каналу з відповідними правами. До того ж, у групах за замовчуванням діє режим конфіденційності – бот бачить лише команди або згадки. Крім цього, бот не має доступу до повідомлень, надісланих до його додавання, що унеможливлює повноцінний аналіз архіву [23].

Для обходу цих обмежень використовується TDLib, які надають більше можливостей у аналізі даних.

2.3.2 Використання TDLib для аналізу даних

TDLib – це кросплатформна, повнофункціональна бібліотека, яка була створена для спрощення розробки власних клієнтів та застосунків, що використовують інфраструктуру Telegram [26].

Вона бере на себе всі аспекти мережевої взаємодії, шифрування та локального зберігання даних, що дозволяє розробникам зосередитися на створенні інтерфейсу та функціональних можливостей додатків.

Серед переваг TDLib – повна підтримка можливостей Telegram, включно з мультимедійним контентом, повідомленнями та керуванням чатами [27]. Як повноцінний клієнт, вона забезпечує стабільну роботу з мережею, зберіганням і синхронізацією даних, відкриваючи широкі можливості для аналізу інформації в месенджері [24].

Ключовими функціями TDLib є обробка текстових повідомлень, мультимедійного контенту, що дозволяє створювати системи автоматичного вилучення даних з каналів, груп, чатів користувача, та подальший аналіз цих даних [25].

2.3.3 Використання Telethon для аналізу даних

Telethon – це асинхронна Python-бібліотека, яка спрощує написання програм, які можуть взаємодіяти з Telegram. Серед функціональних можливостей бібліотеки – доступ до історії чатів, включно з каналами, груповими розмовами та приватними діалогами. Вона дозволяє надсилати як текстові повідомлення, так і мультимедійні файли, здійснювати завантаження та збереження контенту.

Telethon також підтримує моніторинг нових повідомлень у реальному часі, а також забезпечує фільтрацію, пошук і подальший аналіз отриманої інформації [45].

Telethon також підтримує прослуховування подій, зокрема нових повідомлень, реакцій, змін учасників групи тощо, що робить її зручною для створення систем моніторингу або автоматичної модерації [45].

2.3.4 Використання Pyrogram для аналізу даних

Pyrogram – це сучасна, асинхронна структура MTProto API, яка дозволяє створювати як звичайні користувацькі, так і бот-акаунти, забезпечуючи повний доступ до функціоналу Telegram. Pyrogram надає зручні засоби для надсилання текстових повідомлень, медіа, документів, геолокацій, контактів тощо. Також бібліотека дозволяє реалізувати обробку подій, що дає змогу реагувати на дії користувачів – наприклад, створювати ботів модерації або системи моніторингу в реальному часі [46].

3 ДОСЛІДЖЕННЯ ТА АНАЛІЗ ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ ТА ТЕХНОЛОГІЇ OSINT ДЛЯ ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ

Для ефективного виявлення шкідливого контенту дедалі активніше застосовуються інструменти штучного інтелекту, зокрема методи обробки природної мови (NLP), глибокого навчання (LSTM, CNN, трансформери як-от BERT, RoBERTa), а також технології OSINT для збору та аналізу відкритих даних. Розглянемо детальніше ці інструменти.

3.1 Застосування методів NLP для виявлення небезпечного контенту

Обробка природної мови (Natural Language Processing, NLP) є основою підходів до автоматичного аналізу текстових даних у соцмережах. NLP-техніки перетворюють неструктурований текст у форму, придатну для машинного аналізу, і дозволяють виявляти в ньому ознаки образливості або ворожості [28]. У межах NLP модель Bag-of-Words (представлення тексту у вигляді «мішка слів») стала одним із перших векторних підходів до представлення текстових даних для подальшої обробки. Однак такі підходи ігнорують порядок слів і контекст, що може призводити до хибних спрацювань: окремі слова можуть мати різний зміст у різних ситуаціях [29]. Дослідження [29] продемонструвало, що врахування послідовностей слів шляхом використання біґрам та триґрам (n-грам моделей) дозволяє покращити якість класифікації тексту, зокрема у задачах виявлення мови ненависті. Такий підхід частково враховує локальний контекст слів, що є перевагою порівняно з класичною Bag-of-Words моделлю, яка оперує лише окремими лексемами без урахування їхнього порядку. Як зазначено в дослідженні [30], інженерне опрацювання ознак – зокрема, вибір відповідних текстових представлень, використання n-грам, методів відбору ознак та зниження розмірності – справляє значний вплив на точність автоматичного виявлення образливого контенту. Тому різні способи представлення тексту і відбору ознак помітно змінюють результати класифікації образливих коментарів [30]. Крім статистичних ознак, у низці досліджень активно застосовувалися

лексичні ресурси, зокрема списки нецензурної лексики, словники образливих висловів, а також експертно сформульовані правила для виявлення погроз або ознак цькування [29].

Попри успіх окремих таких рішень, класичні підходи мають обмеження щодо врахування багатозначності слів, сарказму, контекстних нюансів і швидко розвиваючогося Інтернет-сленгу. Тому нині основний акцент робиться на більш гнучких та потужних методах штучного інтелекту – передусім глибокого навчання – які можуть автоматично навчатися розпізнавати складні шаблони шкідливого мовлення без ручного конструювання правил [31].

3.2 Глибоке навчання у задачах розпізнавання шкідливих текстових повідомлень

3.2.1 Рекурентні та згорткові нейромережі (RNN, LSTM, CNN)

Методи глибинного навчання наразі домінують у задачі виявлення шкідливого контенту, оскільки вони здатні автоматично виділяти складні мовні закономірності. Широко використовуються моделі на основі рекурентних нейронних мереж (RNN), які пристосовані обробки послідовних даних, таких як текст. Зокрема, мережі довгої короткочасної пам’яті (Long Short-Term Memory, LSTM) – різновид RNN – здатні запам’ятовувати довготривалі залежності у тексті та враховувати широкі контекстні взаємозв’язки слів [30]. Це дозволяє їм уловлювати смисловий контекст образливого висловлювання, навіть якщо образа виражена не прямо, а опосередковано.

Спрощеним різновидом рекурентної мережі є блоки з GRU (Gated Recurrent Unit), які мають схожу архітектуру з меншим числом параметрів. Такі GRU- та LSTM-модулі часто закладаються в ядро сучасних моделей для аналізу тексту поряд із згортковими шарами [32].

Згорткові нейронні мережі (Convolutional Neural Networks, CNN) теж виявилися ефективними для класифікації тексту. Спершу CNN здобули популярність у комп’ютерному зорі, але в обробці мови вони добре виокремлюють ключові локальні шаблони – наприклад, характерні словосполучення або лайливі фрази, що часто вживаються в мові ненависті [33].

Для підвищення ефективності нерідко поєднуються рекурентні та згорткові підходи в одній архітектурі. Наприклад, модель C-BiLSTM (Convolutional Bi-Directional LSTM) – це архітектура глибокого навчання, яка поєднує переваги згорткових шарів (CNN) для виявлення локальних ознак у тексті з можливістю двонаправлених LSTM (BLSTM) вивчати послідовні патерни як у прямому, так і у зворотному напрямках. Згортковий шар генерує векторні представлення фраз, які потім аналізуються BLSTM для моделювання контексту в обох напрямках, а кінцеве рішення про наявність неприйнятного контенту приймається через повнозв'язний класифікаційний шар. Така гібридна архітектура водночас захоплює локальні ознаки тексту (через CNN) і глобальний контекст висловлювання (через двонаправлену LSTM). Моделі цього типу навчаються до кінця без ручної розмітки ознак і продемонстрували вищу точність порівняно з класичними методами машинного навчання. Зокрема, C-BiLSTM суттєво перевершила за якістю патерн-орієнтовані підходи та моделі з ручними ознаками під час виявлення недоречних запитів [31].

3.2.2 Трансформерні моделі (BERT, RoBERTa)

Останнім поколінням моделей NLP, що встановили нові стандарти якості в задачах класифікації тексту, є архітектури на основі трансформерів. Відмінність трансформерних моделей від RNN полягає у здатності обробляти всю послідовність слів паралельно, що забезпечує швидше навчання та обробку довгих послідовностей. Ключовою особливістю трансформерів є механізм самоуваги , який дозволяє моделі враховувати контекст з обох сторін для кожного слова [34, 35]. Яскравим прикладом є модель BERT представлена компанією Google у 2018 році. Інновація BERT полягала в двонаправленому аналізі: модель «розуміє» значення слова з урахуванням повного оточення в реченні [33]. Це виявилося надзвичайно важливим для розпізнавання тонких і завуальованих проявів ненависті, де значення фрази сильно залежить від контексту. Завдяки здатності глибоко враховувати контекст, трансформерні архітектури, зокрема BERT і його покращена версія RoBERTa, продемонстрували високу ефективність у виявленні мови ненависті та токсичного контенту.

На практиці сучасні рішення часто використовують попередньо навчені трансформерні моделі як основу класифікатора. Наприклад, у дослідженні [32] для задачі виявлення шкідливого контенту було протестовано різні глибинні моделі з різними типами векторних представлень слів, серед яких контекстні трансформерні представлення BERT і RoBERTa забезпечили одні з найкращих результатів, особливо при поєднанні з BiLSTM-архітектурою. Комбінація трансформерних ознак із традиційними продемонструвала перевагу трансформерів у розумінні семантики тексту [32]. В цілому, трансформерні підходи наразі задають найвищий рівень якості у завданні автоматичної модерації контенту, тому інтеграція моделей на кшталт BERT/RoBERTa стала стандартною практикою при побудові систем фільтрації ненависних висловлювань.

3.2.3 Моделі для багатомовних і кодозмішаних текстів

У сучасному інформаційному середовищі кодозмішаний текст – тобто повідомлення, що містять елементи кількох мов у межах одного висловлювання, слова чи речення – становить серйозний виклик для систем автоматичного мовного аналізу. Дослідження показують, що подібні мовні конструкції є типовими для спілкування у багатомовних онлайн-спільнотах і масово представлені в соціальних мережах. Змішані мовні висловлювання часто включають внутрішньореченнєве (всередині речення), внутрішньослівне (частини одного слова написані різними мовами) та міжреченнєве (різні речення різними мовами). Це створює складнощі для систем розпізнавання мов, оскільки більшість традиційних моделей працюють лише з цілими реченнями або текстами. Такий підхід не підходить у ситуаціях, коли мовне перемикання відбувається на рівні окремих слів або навіть морфем. Окрім того, моделі, побудовані на словниках або фіксованих правилах, часто не здатні впоратися зі словами, що мають запозичення, нестандартне написання чи фонетичну транслітерацію [36]. Для вирішення цієї проблеми застосовуються мультимовні моделі, які здатні обробляти текст на кількох мовах одночасно.

Мультимовні трансформерні моделі, такі як mBERT, XLM і XLM-R, становлять основу сучасних підходів до обробки текстів кількома мовами, особливо у випадках, коли йдеться про мови з обмеженими наборами даних. Модель mBERT (Multilingual BERT) – це багатомовна версія архітектури BERT, яка була попередньо навчена на текстах 104 мов із використанням маскованого мовного моделювання та передбачення наступного речення. Вона не потребує ручної розмітки та дозволяє вирішувати широкий спектр лінгвістичних завдань. Модель XLM (Cross-Lingual Language Model) – це покращена багатомовна модель, яка навчається не лише на текстах окремих мов, як це було в mBERT, а й на паралельних реченнях різними мовами. Завдяки цьому вона краще «розуміє» зв’язки між мовами. Вона також використовує спеціальне кодування слів, що дозволяє ефективніше працювати з текстами різних мов, знаходити спільне між ними та покращувати точність при розпізнаванні змісту. XLM краще справляється з переносом знань з однієї мови на іншу та показує хороші результати в багатомовних NLP-завданнях. XLM-R (XLM-RoBERTa) є вдосконаленою версією попередніх моделей, навченою на великому корпусі з понад 2,5 ТБ текстів 100 мов. Вона продемонструвала високу точність у завданнях класифікації, розпізнавання іменованих сутностей та відповідей на запитання. Вона також може добре обробляти текст новою мовою, якої не було у тренувальних даних [37].

3.3 Моделі штучного інтелекту для виявлення шкідливого контенту у мультимедійних даних

Окрім текстових повідомлень, важливо аналізувати також мультимедійні файли, такі як зображення, відео та, зокрема, меми. В соціальний мережах, візуальна інформація часто використовується як носій прихованих або завуальованих форм агресії, мови ворожнечі чи дискримінації. Особливо використовується мем-контент, який зазвичай поєднує зображення та короткий текст, що ускладнює виявлення шкідливого сенсу без комплексного аналізу.

Одним із перспективних підходів до виявлення шкідливого контенту у зображеннях є застосування мультимодальних багатозадачних нейронних мереж,

які поєднують аналіз тексту та візуального вмісту. У рамках завдання автоматичного виявлення мізогінії в мультимедійних ЗМІ було розглянуто дві принципово різні архітектурні схеми мультимодальних моделей: одно-потокова трансформерна архітектура та архітектура подвійної башти [38].

Одно-потокова архітектура (single-flow) базується на уніфікованій моделі, що обробляє текстову та візуальну інформацію як єдиний вхід. Спочатку текст і зображення кодуються окремо – текст за допомогою попередньо натренованої моделі BERT, а зображення за допомогою моделей на кшталт ResNet-152 чи EfficientNet. Після цього обидва типи ознак об'єднуються (конкатенуються) у спільний векторний простір і подаються на вхід трансформера. Основною перевагою цього підходу є здатність моделі відображати міжмодальні зв’язки у спільному семантичному просторі, що дозволяє виявляти глибинні кореляції між текстом і зображенням. Водночас ефективність такого підходу значною мірою залежить від попереднього навчання на великих мультимодальних корпусах даних. Альтернативою виступає архітектура подвійної башти, у якій окремі гілки моделі відповідають за аналіз тексту та зображення, що об’єднуються лише на фінальному етапі. Такий підхід виявився ефективним у випадках обмеженого обсягу навчальних даних [38].

Архітектура подвійної башти (double-tower), навпаки, передбачає паралельну обробку зображення та тексту двома окремими гілками. Для кожного типу даних створюється окрема модель (наприклад, ResNet-18 для зображень і Bertweet для тексту), результати якої поєднуються на виході. Хоча цей підхід поступається single-flow за глибиною міжмодальної інтеграції, він має низку переваг: меншу залежність від обсягу навчальних даних, кращу стійкість до перенавчання та більшу гнучкість при обмежених обчислювальних ресурсах. Це робить його придатним для задач, де обсяг навчальної вибірки є невеликим [38].

Для підвищення точності було також реалізовано ансамблювання моделей, яке поєднує прогнози обох архітектур шляхом обчислення зваженого середнього. Такий підхід дозволяє ефективно використовувати сильні сторони кожної з моделей, компенсуючи їхні недоліки [38].

3.4 OSINT як інструмент виявлення шкідливого контенту

3.4.1 Визначення поняття OSINT у контексті боротьби з шкідливим контентом

Open Source Intelligence (OSINT) – це процес збору, обробки, аналізу та інтерпретації інформації, яка отримується виключно з відкритих джерел. До таких джерел належать цифрові публікації, блоги, новини, відеохостинги, відкриті реєстри, соціальні мережі, онлайн-форуми, коментарі, технічні метадані, а також інші легальні та публічно доступні ресурси [39].

У сфері інформаційної безпеки OSINT є окремим напрямом розвідки, що використовується як державними структурами, так і приватними компаніями для виявлення інформаційних загроз, оцінки репутаційних ризиків, протидії злочинним або ворожим інформаційним впливам [39]. Одним із ключових напрямів використання OSINT сьогодні є боротьба зі шкідливим контентом, під яким розуміють публікації, що сприяють поширенню дезінформації, мови ворожнечі, закликів до насильства, самозашкодження, розпалювання ворожнечі, дискримінації або інших форм контенту, здатного завдати шкоди суспільству, окремим групам чи особам [39].

Інтеграція OSINT із технологіями штучного інтелекту відкриває нові можливості для аналізу шкідливого контенту. З одного боку, саме алгоритми штучного інтелекту є основним інструментом обробки інформації, зібраної OSINT-засобами. За допомогою ШІ застосовуються різноманітні підходи до аналізу даних, зокрема аналіз змісту, відстеження поведінки користувачів, мережевого аналізу. Аналіз змісту повідомлень дозволяє ідентифікувати ознаки дезінформації, маніпуляцій або емоційного впливу шляхом виявлення логічних суперечностей, неточностей та характерних стилістичних маркерів. Окрему увагу приділяють відстеженню активності користувачів: час публікацій, частота повідомлень, а також поведінкові характеристики можуть слугувати індикаторами потенційної загрози або організованої кампанії. Крім того, використовується аналіз зв’язків між окремими обліковими записами, сторінками або групами, який дозволяє виявити приховані мережі впливу або

скоординовані дії, спрямовані на поширення шкідливого контенту. У поєднанні ці підходи забезпечують комплексний огляд інформаційного середовища [39].

З іншого боку, дані, зібрані з відкритих джерел, використовуються як навчальні вибірки для побудови моделей машинного навчання. Наприклад, текстові повідомлення, зібрані за допомогою OSINT-інструментів, можуть бути класифіковані за рівнем агресивності або за типом загрози – мова ворожнечі, фейкова новина, погроза тощо. У процесі навчання такі набори даних використовуються для формування класифікаційних моделей, які надалі можуть самостійно ідентифікувати подібні патерни в новому контенті [39].

У поєднанні ці дві технології – OSINT, як джерело, та AI, як інструмент аналізу – формують цілісну систему моніторингу, яка дозволяє не лише виявляти шкідливий контент, а й прогнозувати інформаційні загрози та моделювати сценарії їх розвитку.

3.4.2 Загальні OSINT-інструменти для збору та обробки даних

Для ефективного збору та аналізу відкритої інформації в Інтернеті розроблено цілу низку спеціалізованих OSINT-інструментів. Вони дають змогу автоматизувати процес моніторингу, фільтрації та візуалізації онлайн-контенту, що потенційно може становити загрозу, зокрема у вигляді шкідливих повідомлень, дезінформації, радикальних закликів або мови ворожнечі.

Одним із ключових інструментів у сфері OSINT є Maltego – програмна платформа для візуального аналізу даних, яка дозволяє виявляти зв’язки між об’єктами, такими як IP-адреси, домени, облікові записи, особи та організації. Це складний інструмент, який здійснює пошук в декількох джерелах даних і представляє результати в графічній формі [40, 39].

Ще одним популярним засобом є TheHarvester – це інструмент OSINT-розвідки, призначений для збору електронних адрес, піддоменів, імен та іншої інформації з відкритих джерел. Він використовує пошукові системи, такі як Google, Bing, Yahoo, а також ресурси типу Shodan та DNSdumpster для збору даних про компанії, організації або цільові домени. Інструмент є корисним як для фахівців із кібербезпеки, так і для OSINT-аналітиків при проведенні

поверхневого розвідувального збору перед тестуванням на проникнення або в рамках аналізу інформаційних слідів у соціальних медіа [41, 39].

Для обробки та аналізу мультимедійного контенту в рамках OSINT-розвідки широко застосовується ExifTool – спеціалізоване програмне забезпечення для вилучення, перегляду та аналізу метаданих із файлів. ExifTool підтримує велику кількість форматів – зокрема, JPEG, PNG, TIFF, MP4, PDF, а також багато форматів аудіо- та відеофайлів. Програма дозволяє отримати детальну інформацію про час і умови створення файлу, пристрій, геолокацію (у разі її наявності), версію програмного забезпечення, яке використовувалось при редагуванні, та інші технічні параметри. Завдяки цим можливостям ExifTool активно використовується для встановлення автентичності зображень або відео, перевірки їхнього джерела та виявлення можливих маніпуляцій. У межах OSINT така інформація є цінною при вивченні фейкового або шкідливого контенту, особливо у випадках розслідувань, пов’язаних з пропагандою, дезінформацією або кіберзлочинами [42, 39].

Окрему категорію OSINT-інструментів становлять ті, що допомагають встановлювати зв’язки між цифровими ідентичностями користувачів. Один із таких Whatsmyname.app – це OSINT-інструмент, який дає змогу автоматично перевіряти наявність певного імені користувача на багатьох веб-сайтах одночасно. Він має велику базу сайтів, де можна перевірити, чи використовується певне ім’я користувача. Такий підхід дозволяє встановити зв’язки між обліковими записами однієї особи в різних платформах. Чим рідкісніше й унікальніше ім’я користувача, тим вищі шанси на точне зіставлення. Завдяки цьому дослідники можуть швидко виявити, на яких онлайн-платформах використовується той чи інший нікнейм, що робить цей інструмент надзвичайно корисним для цифрових розслідувань та моніторингу цифрової активності конкретного користувача [43, 39].

Незважаючи на те, що більшість описаних інструментів не є спеціалізованими засобами для автоматичного виявлення шкідливого контенту, вони відіграють важливу роль у його дослідженні. Наприклад, Maltego дозволяє

виявляти зв’язки між цифровими об’єктами, що може допомогти у виявленні скоординованих інформаційних кампаній чи мереж поширення дезінформації. TheHarvester надає можливість швидко отримати контактні та технічні дані, пов’язані з потенційно небезпечними джерелами контенту. ExifTool є корисним при перевірці метаданих зображень і відео, що дозволяє встановити їхню автентичність і джерело – наприклад, при розслідуванні фейкових або маніпулятивних матеріалів. А інструмент WhatsMyName дає змогу встановити, на яких платформах використовується певне ім’я користувача, що сприяє ідентифікації активності суб’єктів, пов’язаних з поширенням шкідливого контенту. У комплексі ці засоби суттєво підсилюють можливості OSINT-аналітики в контексті боротьби з інформаційними загрозами.

3.4.3 Спеціалізовані OSINT-системи для виявлення шкідливого контенту

Окрему категорію OSINT-інструментів становлять ті, що призначені безпосередньо для виявлення шкідливого контенту, зокрема дезінформації, мови ворожнечі, закликів до насильства або психологічного тиску. Ці інструменти часто інтегрують технології штучного інтелекту, включаючи обробку природної мови, семантичний аналіз, виявлення аномалій та автоматичну класифікацію повідомлень.

Однією з показових реалізацій застосування OSINT у поєднанні з технологіями штучного інтелекту є система, розроблена для виявлення кіберінцидентів та загроз на основі аналізу публікацій у Twitter. Цей підхід був запропонований дослідниками у статті [44], де детально описано архітектуру багаторівневої аналітичної системи, здатної автоматично ідентифікувати загрози, пов’язані з кібербезпекою.

Рішення складається з кількох етапів. На першому рівні за допомогою OSINT-інструменту Twint здійснюється автоматизований збір твітів за заданими ключовими словами. Далі йде модуль фільтрації, що дозволяє відсіювати нерелевантний контент. На наступному етапі використовується модель на основі рекурентної нейронної мережі для класифікації повідомлень як таких, що стосуються кіберподій, або нейтральних [44].

Після система виконує розпізнавання іменованих сутностей, що є методом обробки природної мови і дозволяє автоматично виділяти з текстів назви зловмисників, шкідливого програмного забезпечення, компаній або інших ключових об'єктів кіберподій. Отримані сутності використовуються для побудови мережі співзгадуваності (Co-occurrence network), яка відображає зв’язки між ними та дозволяє згрупувати повідомлення навколо окремих подій [44].

На останньому етапі проводиться оцінка достовірності інформації за допомогою двох спеціалізованих метрик: Diffusion Index, що вимірює рівень поширення твітів, та Spam Index, який аналізує їх правдоподібність і ризик автоматизованої генерації [44].

4 ПЕРСПЕКТИВИ ТА РЕКОМЕНДАЦІЇ ЩОДО РОЗВИТКУ ТЕХНОЛОГІЙ ДЛЯ МОНІТОРИНГУ TELEGRAM-КАНАЛІВ

4.1 Обґрунтування потреби в програмному рішенні

Зараз, в Україні Telegram став одним із найпопулярніших майданчиків для обміну інформацією. В умовах повномасштабної війни, Telegram виконує дві ролі: з одного боку він є джерелом для отримання новин, а з іншого його використовують для поширення фейків, пропаганди та небезпечного контенту.

Ручний моніторинг Telegram-каналів, групи є нереальним завданням, враховуючи колосальну кількість інформації щодня. Крім того, багато шкідливих повідомлень можуть бути замасковані під звичайні, тобто не мати явних заборонених слів, але визначати ворожий контекст. Усе це створює потребу в інструменті, який допоміг би автоматично виявляти потенційно небезпечний контент, аналізуючи не лише слова, а й загальний контекст повідомлень.

Саме тому виникла ідея створити програмне рішення, яке може:

- підключатись до Telegram і аналізувати публічні канали за допомогою OSINT;
- обробляти повідомлення, автоматично визначати мову та робити базову обробку тексту для подальшого аналізу;
- використовувати штучний інтелект для аналізу змісту та класифікації повідомлень за рівнем потенційної загрози (мова ворожнечі, ідеологічна пропаганда, контент, що може зашкодити психічному здоров’ю тощо);
- створювати звіти, які можуть бути корисні для аналітиків, дослідників або служб інформаційної безпеки.

Розроблений застосунок буде корисним для різних категорій користувачів. Фахівці з інформаційної безпеки отримають інструмент для моніторингу Telegram без необхідності аналізувати повідомлення вручну. Журналісти й фактчекери зможуть ефективніше перевіряти підозрілі канали. Дослідники отримають джерело якісно структурованих даних для аналізу. І навіть звичайні

користувачі зможуть використовувати застосунок для власного інформаційного захисту.

4.2 Обґрунтування використаних технологій та огляд розробленого рішення

Для реалізації програмного засобу було обрано мову програмування Python, яка на сьогодні є однією з найпопулярніших мов у сфері штучного інтелекту та аналізу даних [47]. Python відзначається простим синтаксисом і великою кількістю готових бібліотек, тому це спрощує розробку та інтеграцію необхідних компонентів.

Для збору повідомлень із Telegram-каналів було застосовано бібліотеку Telethon, що є асинхронним клієнтом Telegram API на Python. Telethon надає зручний інтерфейс для авторизації користувача та отримання даних з чатів, виконуючи більшість низькорівневих операцій замість розробника [45]. Вибір Telethon обумовлений його поширеністю в OSINT-дослідженнях і здатністю працювати безпосередньо від імені користувача Telegram [48]. Важливо, що Telethon в поєднанні з бібліотекою asyncio побудований на принципах асинхронності, що дозволяє ефективніше отримувати великі обсяги даних, оскільки інші операції можуть виконуватися паралельно із очікуванням мережевих відповідей, що особливо важливо при моніторингу кількох каналів Telegram.

Для автоматичної класифікації контенту було обрано трансформерну модель «MoritzLaurer/mDeBERTa-v3-base-mnli-xnli», доступну через бібліотеку Hugging Face Transformers. Дана багатомовна модель була попередньо навчена на задачі природної мовної інференції (NLI) для 100 мов, зокрема на англомовному датасеті MNLI та багатомовному XNLI [49]. Це означає, що модель здатна розуміти семантичні зв’язки і робити zero-shot класифікацію, тобто визначати належність тексту до нових категорій без додаткового донавчання на спеціалізованих вибірках. Вибір саме цієї моделі пояснюється її високими показниками: на кінець 2021 року вона була однією з найкращих серед базових багатомовних моделей [49]. Таким чином, використання попередньо навченого

трансформера дозволяє досягти високої якості класифікації шкідливого контенту без потреби збирати великий корпус даних для навчання з нуля.

Застосований підхід нульової класифікації (zero-shot) зумовлений тим, що він дає змогу моделі класифікувати повідомлення за наперед заданими категоріями без прямого навчання на прикладах саме цих категорій. Модель, навчена на NLI, фактично оцінює правдивість гіпотези на кшталт “це повідомлення містить <категорія>” для кожної категорії і таким чином вирішує, чи відповідає текст даній тематиці. Всього в програмі використовуються 4 категорії, які вказують на відповідний шкідливий контент:

- "Hate speech or harassment",
- "Propaganda or ideological harm",
- "Exploitation or abuse",
- "Self-harm or suicide".

Та одна категорія “Safe”, яка класифікує не шкідливий контент. Така методика активно використовується у завданнях модерації контенту, бо допомагає виявляти неприйнятний або шкідливий контент на різних платформах, дозволяючи фіксувати нові типи порушень без наявності явних прикладів для навчання [50]. Для Telegram це особливо важливо, оскільки можуть з’являтися нові форми дезінформації чи мови ворожнечі, які важко передбачити наперед традиційними методами навчання.

При розробці програмного забезпечення, викликом стало те, що сирі повідомлення можуть мати HTML-теги, посилання, емодзі, спецсимволи, або бути надто короткими, що унеможливлює аналіз трансформерною моделлю. Рішенням стала попередня обробка даних, яка видаляє HTML-розмітку, URL-адреси, непотрібні символи, приводить текст до єдиного регістру. Окрім цього, відфільтровує повідомлення, які не несуть лінгвістичної інформації (наприклад, складаються лише з цифр чи символів, або містять менш, ніж 3 символи). Попереднє очищення і нормалізація тексту підвищують якість і достовірність подальшої класифікації, оскільки модель отримує на вхід стандартизовані дані без зайвих перешкод [51].

Архітектура програмного забезпечення побудована модульно, із чітким розподілом функцій між компонентами. Виділено окремі модулі для взаємодії з Telegram (telegram_module), для попередньої обробки даних (preprocessing), для класифікації тексту (classifier), генерації звіту (report_generator), а також утилітні функції (utils).

Модульна структура програми робить її гнучкою і зручною для подальшого розвитку. Окремі компоненти, такі як класифікатор або генератор звітів, працюють незалежно один від одного, що дозволяє використовувати їх не лише для аналізу Telegram, а й для інших платформ. Для цього достатньо лише надати їм дані у правильному форматі. Такий підхід відкриває можливість без особливих труднощів підключати нові джерела інформації, наприклад, Twitter чи інші соціальні мережі: достатньо розробити новий модуль збору даних, тоді як класифікація й створення звітів залишаються незмінними.

Архітектура також підтримує одночасний аналіз кількох каналів Telegram. Завдяки асинхронному підходу програма може паралельно обробляти дані з різних джерел, зберігаючи результати окремо для кожного каналу. Логіка класифікації легко налаштовується під нові завдання: додати нову категорію шкідливого контенту можна просто, змінивши перелік міток у класифікаторі, без необхідності переписувати всю систему. Усе це робить розроблену програму гнучкою, адаптивною та готовою до нових викликів.

Під час розробки було використано низку сторонніх бібліотек, які допомогли розширити можливості застосунку та спростити реалізацію окремих його частин. Таким чином, бібліотека Transformers від Hugging Face стала основним інструментом для роботи з мовною моделлю: завдяки зручному інтерфейсу pipeline вдалося легко інтегрувати передові методи обробки природної мови та організувати процес zero-shot класифікації повідомлень. Для візуалізації результатів аналізу в модулі генерації звітів було обрано matplotlib – ця бібліотека дозволила швидко створювати наочні графіки, зокрема діаграми розподілу повідомлень за категоріями.

Крім того, активно застосовуються стандартні бібліотеки Python – json, os та re. Вони відповідають за збереження даних у форматі JSON, взаємодію з файловою системою та очищення тексту за допомогою регулярних виразів.

Щоб зробити роботу програми швидшою та ефективнішою при обробці великих обсягів даних, було впроваджено механізм кешування результатів класифікації. Після першої обробки повідомлення результат зберігається у локальний кеш, тож, якщо той самий текст трапляється знову – у тому ж чи іншому каналі або навіть під час наступного запуску програми – система просто підтягує вже готову відповідь замість повторного аналізу.

Такий підхід дозволяє значно економити ресурси: кешування допомагає уникнути зайвих розрахунків і помітно скорочує час обробки. Це особливо важливо у випадку використання трансформерних моделей, адже кожен запуск такої моделі є доволі ресурсомістким. Завдяки кешу навантаження на систему знижується, а результати аналізу можна отримувати набагато швидше навіть при роботі з великими вибірками повідомлень.

Після реалізації програмного забезпечення було проведено тестування його функціональності для підтвердження працездатності запропонованого рішення. Програмне забезпечення встановлюється та запускається відповідно до інструкції, яка наведена у додатку Б.

Після запуску файлу main.py програма ініціює процес авторизації в Telegram через API. Користувач вводить номер телефону, отримує код підтвердження, за потреби у користувача також може запитати двохфакторний пароль. Після успішної автентифікації виконується вхід. На рис. В.1 в додатку В наведено приклад авторизації.

Після авторизації програма автоматично отримує список каналів, доступних користувачеві, та виводить їх на екран із порядковими номерами. Користувач вибирає номер потрібного каналу для збору повідомлень. Після чого користувачеві пропонується вказати кількість повідомлень, які треба завантажити з обраного каналу. Перед початком класифікації, програма запитує

підтвердження на початок аналізу. Приклад вибору каналів та кількості повідомлень наведено на рис. В.2 в додатку В.

Зібрані повідомлення проходять етап попередньої обробки. Після цього здійснюється класифікація повідомлень за категоріями шкідливого контенту. На основі класифікованих повідомлень формуються підсумкові звіти. На рис. В.3 в додатку В показано приклад, який описує процес класифікації з індикатором виконання, повідомленням про генерацію звітів та прикладом короткого підсумкового звіту із зазначенням кількості повідомлень у кожній категорії.

Згенеровані звіти зберігаються за шляхом “/reports/[назва_каналу]/”. Кількість згенерованих звітів для одного каналу може варіюватися від 5 (один за категорією та чотири загальні) до 9 (п'ять за категоріями та чотири загальні). На рис. В.4 в додатку В представлено приклад структури зберігання звітів, які були сформовані у ході попередніх прикладів.

На рис. В.5 в додатку В показаний приклад, вмісту одного з json файлів звіту за категоріями (category_Exploitation_or_abuse.json). На рис. В.6 в додатку В показаний приклад діаграми розподілу повідомлень за категоріями шкідливого контенту (chart.png). Файл classification_summary.txt містить інформацію про кількість повідомлень у кожній категорії, їхнє відсоткове співвідношення, а також середню довжину повідомлень у відповідній категорії. Файл top_toxic_messages.txt містить топ-5 найбільш токсичних і шкідливих повідомлень. У файлі top_uncertain_examples.txt зберігаються топ повідомлень із найвищим рівнем невизначеності, на які слід звернути особливу увагу.

4.3 Напрями подальшого розвитку

Розроблене програмне забезпечення для виявлення шкідливого контенту в Telegram-каналах демонструє свою ефективність, проте існує низка можливостей для його подальшого вдосконалення та розширення функціональності. Перспективні напрями розвитку охоплюють підвищення точності класифікації, масштабування системи на інші платформи, розширення аналітичних

можливостей для OSINT та підтримку мультимедійного контенту. Нижче детально розглянуто кожен з цих напрямів.

Одним із ключових напрямів вдосконалення є підвищення точності класифікації повідомлень. У поточній реалізації використано попередньо навчений багатомовний трансформер у режимі zero-shot класифікації без додаткового донавчання на прикладах конкретних категорій. Хоча такий підхід показав практичність і непогані результати, ще більшої точності можна досягти шляхом тонкого налаштування моделі (fine-tuning) на спеціалізованому наборі даних. Для цього доцільно зібрати корпус повідомлень з Telegram-каналів, вручну анотований за категоріями шкідливого контенту, і донавчити на ньому сучасну трансформерну модель. У результаті модель, адаптована до специфіки Telegram-контенту, зможе значно підвищити надійність і точність системи класифікації шкідливого контенту.

Ще один важливий напрям подальшого розвитку полягає у масштабуванні програмного забезпечення на інші платформи. Зараз система орієнтована на аналіз контенту Telegram, але потенційно її можна застосувати й до інших соціальних мереж та месенджерів (Twitter, Reddit, Facebook тощо), які також є джерелами шкідливого контенту. Для цього достатньо реалізувати окремі модулі збору інформації через API інших платформ: наприклад, інтегрувати API Twitter для отримання твітів або API Reddit для збору коментарів і постів. Після збору даних з нового джерела вони будуть проходити через ті самі етапи попередньої обробки та класифікації, за умови, що мовна модель була належним чином адаптована до особливостей відповідного середовища. Враховуючи відмінності у форматі повідомлень, можливо, знадобиться додаткове тонке налаштування моделі під кожну платформу.

Окремо слід виділити розширення функціональності для OSINT-дослідників. Поточна версія програмного забезпечення генерує підсумкові звіти щодо виявленого шкідливого контенту, проте аналітикам можуть знадобитися глибші й різноманітніші інструменти аналізу. Тому важливим удосконаленням може стати впровадження додаткових форматів звітності таких як PDF або

HTML, що міститимуть детальну статистику, графіки динаміки появи шкідливих повідомлень та приклади характерного контенту. Також доцільно реалізувати автоматичну генерацію аналітичних висновків, зокрема виявлення тематичних трендів, періодів активності шкідливого контенту та нетипових сплесків, що можуть свідчити про інформаційні атаки. Інтеграція з існуючими системами моніторингу через API дозволить відстежувати появу небезпечного контенту в реальному часі та оперативно інформувати аналітиків. У результаті розширення функцій перетворить систему з утиліти для класифікації на повноцінний аналітичний інструмент для розвідки відкритих даних.

Ще одним перспективним напрямом розвитку є доповнення класифікації обробкою медіаконтенту – зображень, відео та аудіоповідомлень. Оскільки шкідливий контент у Telegram поширюється не лише текстовими повідомленнями, а й через меми, фотографії, відеоролики та голосові записи, текстового аналізу недостатньо для повноцінного моніторингу. Для розширення можливостей системи необхідно інтегрувати методи комп’ютерного бачення та обробки аудіоданих. Такий мультимодальний підхід дозволяє суттєво підвищити ефективність виявлення небезпечної інформації, незважаючи на використання більших обчислювальних ресурсів.

Окрім того, для підвищення зручності користувача в майбутньому доцільно реалізувати графічний інтерфейс. У перспективі розглянуті удосконалення можуть перетворити розроблену систему на універсальний інструмент моніторингу та протидії небезпечному контенту в Інтернеті, здатний відповідати на сучасні виклики інформаційної безпеки.

ВИСНОВКИ

У рамках кваліфікаційної роботи було досягнуто поставлену мету – досліджено існуючих інструментів для моніторингу та аналізу контенту в Telegram, окрім того було проаналізовано можливості застосування штучного інтелекту та технології OSINT для виявлення шкідливого контенту в Telegram-каналах на практиці.

У першому розділі було проведено ґрунтовний аналіз класифікації шкідливого контенту та форм його поширення в інформаційному середовищі , а також існуючих підходів до модерації такого контенту, зокрема ручних та автоматизованих методів. На основі того було виконано порівняльний аналіз ефективності ручних і автоматизованих стратегій модерації контенту, що дозволило обґрунтувати доцільність використання гібридного підходу.

У другому розділі увага була приділена месенджері Telegram, як платформі яка через свою анонімність, відсутність прозорої модерації та гнучкість API, часто використовується для розповсюдження небезпечного контенту. Це дозволило детально оцінити існуючі способи регулювання шкідливого контенту у цьому застосунку та розглянути інструменти Telegram API (зокрема, Telethon, TDLib, Bot API та Pyrogram), що дозволяють ефективно автоматизувати процес збору та аналізу даних, для подальшого дослідження.

У третьому розділі було розглянуто сучасні підходи до виявлення шкідливого контенту із застосуванням штучного інтелекту, включно з методами обробки природної мови (NLP), трансформерними моделями (BERT, RoBERTa, mBERT, XLM-R) та архітектурами глибокого навчання (CNN, RNN, LSTM). Окрім того, було показано роль технології OSINT у доповненні технічних рішень – завдяки здатності виявляти пов’язані джерела інформації, відстежувати взаємозв’язки між акаунтами, каналами та розповсюдженням контенту.

У результаті проведеного аналізу, у четвертому розділі було розроблено інструмент для автоматизованого виявлення небезпечного контенту в Telegram-каналах. Запропоноване рішення поєднує інструменти AI, OSINT та Telegram API

у єдину систему, яка потенційно може бути використана для створення масштабованого програмного забезпечення в сфері інформаційної безпеки. Подальше вдосконалення системи може включати підвищення точності класифікації через донавчання моделі, розширення функціональності для аналізу інших платформ і мультимедійного контенту, а також інтеграцію з аналітичними інструментами для OSINT-дослідників, що дозволить створити більш ефективний інструмент для моніторингу та боротьби з небезпечним контентом.

Таким чином поставленна мета була досягнута. Результати проведенної роботи надають основу для подальшого поєднання технологій OSINT та штучного інтелекту у дослідженнях, спрямованих на захист інформаційного середовища від неприйнятного контенту.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Social media and digital technology use among Indigenous young people in Australia: a literature review / E. S. Rice та ін. *International Journal for Equity in Health*. 2016. Т. 15, № 1. Режим доступу: https://doi.org/10.1186/s12939-016-0366-0 (дата звернення: 15.01.2025).
2. Biggest social media platforms by users 2025| Statista. *Statista*. Режим доступу: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (дата звернення: 15.01.2025).
3. Kennedy T. Indigenous peoples' experiences of harmful content on social media. Sydney : Macquarie University, 2020. 21 с. Режим доступу: https://researchers.mq.edu.au/en/publications/indigenous-peoples-experiences-of-harmful-content-on-social-media (дата звернення: 15.01.2025).
4. Interfax-Ukraine. Ткаченко про телеграм: Я не говорив би про заборону, я говорив би про певне регулювання. *Інтерфакс-Україна*. Режим доступу: https://interfax.com.ua/news/telecom/906774.html (дата звернення: 16.01.2025).
5. IBM. What Is Artificial Intelligence (AI)? | IBM. *IBM - United States*. Режим доступу: https://www.ibm.com/think/topics/artificial-intelligence (дата звернення: 16.01.2025).
6. What is OSINT (Open-Source Intelligence?) | SANS Institute. *Cyber Security Training | SANS Courses, Certifications & Research*. Режим доступу: https://www.sans.org/blog/what-is-open-source-intelligence/ (дата звернення: 16.01.2025).
7. Небезпечний контент - Політика щодо створеного користувачами вмісту Карт Довідка. *Google Help*. Режим доступу: https://support.google.com/contributionpolicy/answer/11409560?hl=uk (дата звернення: 05.02.2025).

8. Шкідливий контент. *Кібер Брама*. Режим доступу: https://stopfraud.gov.ua/cybersecurity-in-education/shkidlyvyj-kontent-i209 (дата звернення: 05.02.2025).
9. Шкідливий контент. Як захистити дітей в інтернеті?. *Stop Sexting - Parents*. Режим доступу: https://stop-sexting.in.ua/adult/ryzyk/shkidlyvyy-kontent/ (дата звернення: 05.02.2025).
10. Banko M., MacKeen B., Ray L. A Unified Taxonomy of Harmful Content. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, м. Online. Stroudsburg, PA, USA, 2020. Режим доступу: https://doi.org/10.18653/v1/2020.alw-1.16 (дата звернення: 05.02.2025).
11. Учасники проектів Вікімедіа. Тероризм – Вікіпедія. *Вікіпедія*. Режим доступу: https://uk.wikipedia.org/wiki/Тероризм (дата звернення: 05.02.2025).
12. Організація Об'єднаних Націй. Terrorism Prevention Branch. *United Nations Office for Drug Control and Crime Prevention*. Режим доступу: https://www.unodc.org/pdf/leaflet_2000-04-30_1.pdf (дата звернення: 05.02.2025).
13. Про боротьбу з тероризмом : Закон України від 20.03.2003 № 638-IV : станом на 9 січ. 2025 р. Режим доступу: https://zakon.rada.gov.ua/laws/show/638-15#Text (дата звернення: 05.02.2025).
14. Gongane V. U., Munot M. V., Anuse A. D. Correction to: Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*. 2022. Т. 12, № 1. Режим доступу: https://doi.org/10.1007/s13278-022-00991-9 (дата звернення: 13.02.2025).
15. Manual Content Analysis: A Beginner's Guide - Insight7 - AI Tool For Interview Analysis & Market Research. *Insight7 - AI Tool For Interview Analysis & Market Research*. Режим доступу: https://insight7.io/manual-content-analysis-a-beginners-guide/ (дата звернення: 13.02.2025).

16. Ruckenstein M., Turunen L. L. M. Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*. 2019. Т. 22, № 6. С. 1026–1042. Режим доступу: https://doi.org/10.1177/1461444819875990 (дата звернення: 13.02.2025).
17. Telegram FAQ. *Telegram*. Режим доступу: https://telegram.org/faq (дата звернення: 19.02.2025).
18. Channels FAQ. *Telegram*. Режим доступу: https://telegram.org/faq_channels (дата звернення: 19.02.2025).
19. Terms of Service. *Telegram*. Режим доступу: https://telegram.org/tos/eu (дата звернення: 19.02.2025).
20. User guidance for the EU Digital Services Act. *Telegram*. Режим доступу: https://telegram.org/tos/eu-dsa (дата звернення: 19.02.2025).
21. Telegram Moderation Overview. *Telegram*. Режим доступу: https://telegram.org/moderation (дата звернення: 20.02.2025).
22. Telegram APIs. *Telegram APIs*. Режим доступу: https://core.telegram.org/api (дата звернення: 20.02.2025).
23. Telegram Bot API. *Telegram APIs*. Режим доступу: https://core.telegram.org/bots/api (дата звернення: 20.02.2025).
24. TDLib: TDLib. *Telegram APIs*. Режим доступу: https://core.telegram.org/tdlib/docs/ (дата звернення: 20.02.2025).
25. Getting started with TDLib. *Telegram APIs*. Режим доступу: https://core.telegram.org/tdlib/getting-started (дата звернення: 20.02.2025).
26. Telegram Database Library. *Telegram APIs*. Режим доступу: https://core.telegram.org/tdlib (дата звернення: 20.02.2025).
27. TDLib – Build Your Own Telegram. *Telegram*. Режим доступу: https://telegram.org/blog/tdlib (дата звернення: 20.02.2025).
28. IBM. What Is NLP (Natural Language Processing)? | IBM. *IBM - United States*. Режим доступу: https://www.ibm.com/think/topics/natural-language-processing (дата звернення: 01.03.2025).

29. Burnap P., Williams M. L. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*. 2015. Т. 7, № 2. С. 223–242. Режим доступу: https://doi.org/10.1002/poi3.85 (дата звернення: 01.03.2025).
30. Habesha@DravidianLangTech: Abusive Comment Detection using Deep Learning Approach / Mesay Gemeda Yigezu та ін. *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. Varna, Bulgaria, 2023. С. 244–249. Режим доступу: https://aclanthology.org/2023.dravidianlangtech-1.36/ (дата звернення: 01.03.2025).
31. Deep learning for detecting inappropriate content in text / H. Yenala та ін. *International Journal of Data Science and Analytics*. 2017. Т. 6, № 4. С. 273–286. Режим доступу: https://doi.org/10.1007/s41060-017-0088-4 (дата звернення: 01.03.2025).
32. StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms. *arXiv.org e-Print archive*. Режим доступу: https://arxiv.org/html/2411.06138v1 (дата звернення: 01.03.2025).
33. Lidoma@DravidianLangTech 2024: Identifying Hate Speech in Telugu Code-Mixed: A BERT Multilingual / Muhammad Zamir та ін. *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. St. Julian's, Malta, 2024. С. 101–106. Режим доступу: https://aclanthology.org/2024.dravidianlangtech-1.16/ (дата звернення: 01.03.2025).
34. Zhaozhen Xu. From RNNs to Transformers. *Baeldung*. Режим доступу: https://www.baeldung.com/cs/rnns-transformers-nlp (дата звернення: 06.03.2025).
35. GeeksforGeeks. Attention vs. Self-Attention in Transformers - GeeksforGeeks. *GeeksforGeeks*. Режим доступу: https://www.geeksforgeeks.org/attention-vs-self-attention-in-transformers/ (дата звернення: 06.03.2025).

36. A Systematic Review on Language Identification of Code-Mixed Text: Techniques, Data Availability, Challenges, and Framework Development / A. F. Hidayatullah та ін. *IEEE Access*. 2022. С. 1. Режим доступу: https://doi.org/10.1109/access.2022.3223703 (дата звернення: 06.03.2025).
37. Cross-lingual Machine Translation: An Analysis Model for Low Resource Languages / G. Bharathi Mohan та ін. *Lecture Notes in Networks and Systems*. Singapore, 2023. С. 81–94. Режим доступу: https://doi.org/10.1007/978-981-99-3963-3_7 (дата звернення: 07.03.2025).
38. Li D., Yi M., He Y. AMS_ADRN at SemEval-2022 Task 5: A Suitable Image-text Multimodal Joint Modeling Method for Multi-task Misogyny Identification. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, м. Seattle, United States. Stroudsburg, PA, USA, 2022. Режим доступу: https://doi.org/10.18653/v1/2022.semeval-1.97 (дата звернення: 07.03.2025).
39. Szymoniak S., Foks K. Open Source Intelligence Opportunities and Challenges: a Review. *Advances in Science and Technology Research Journal*. 2024. Т. 18, № 3. С. 123–139. Режим доступу: https://doi.org/10.12913/22998624/186036 (дата звернення: 07.03.2025).
40. Contributors to Wikimedia projects. Maltego - Wikipedia. *Wikipedia, the free encyclopedia*. Режим доступу: https://en.wikipedia.org/wiki/Maltego (дата звернення: 20.03.2025).
41. GitHub - laramies/theHarvester: E-mails, subdomains and names Harvester - OSINT. *GitHub*. Режим доступу: https://github.com/laramies/theHarvester (дата звернення: 20.03.2025).
42. *ExifTool by Phil Harvey*. Режим доступу: https://exiftool.org/ (дата звернення: 20.03.2025).
43. WhatsMyName Web. *Forensic OSINT: Full Page Web Capture, One Screen at a Time*. Режим доступу: https://www.forensicosint.com/kb/WhatsMyName-Web-ADB3B20924F01_lv3 (дата звернення: 21.03.2025).

44. Dale D., McClanahan K., Li Q. AI-based Cyber Event OSINT via Twitter Data. *2023 International Conference on Computing, Networking and Communications (ICNC)*, м. Honolulu, HI, USA, 20–22 лют. 2023 р. 2023. Режим доступу: https://doi.org/10.1109/icnc57223.2023.10074187 (дата звернення: 21.03.2025).
45. Telethon's Documentation – Telethon 1.37.0 documentation. *Telethon*. Режим доступу: https://docs.telethon.dev/ (дата звернення: 23.03.2025).
46. Welcome to Pyrogram – Pyrogram Documentation. *Pyrogram*. Режим доступу: https://docs.pyrogram.org/ (дата звернення: 23.03.2025).
47. Karl T. 6 Reasons Why Is Python Used for Machine Learning. *New Horizons*. Режим доступу: https://www.newhorizons.com/resources/blog/why-is-python-used-for-machine-learning (дата звернення: 24.03.2025).
48. Telegram OSINT: The Ultimate Guide. *ESPY - Data Enrichment*. Режим доступу: https://espysys.com/blog/telegram-osint-the-ultimate-guide/ (дата звернення: 24.03.2025).
49. MoritzLaurer/mDeBERTa-v3-base-mnli-xnli · Hugging Face. *Hugging Face – The AI community building the future.* Режим доступу: https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli (дата звернення: 24.03.2025).
50. How Zero-Shot Classification Enhances AI Models. *TiDB*. Режим доступу: https://www.pingcap.com/article/how-zero-shot-classification-enhances-ai-models/ (дата звернення: 24.03.2025).
51. Text Preprocessing. *Alooba*. Режим доступу: https://www.alooba.com/skills/concepts/natural-language-processing/text-preprocessing/ (дата звернення: 24.03.2025).

ДОДАТОК А

СТАТИСТИЧНІ ДАНІ БЛОКУВАНЬ ГРУП ТА КАНАЛІВ У TELEGRAM

Рисунок А.1 – Кількість заблокованих груп та каналів модераторами Telegram [21]

Рисунок А.2 – Кількість терористичних спільнот[21]

ДОДАТОК Б

ІНСТРУКЦІЯ З ВСТАНОВЛЕННЯ ТА ВИКОРИСТАННЯ TELEGRAM CONTENT ANALYZER

Б.1 Загальний опис та основні можливості

Telegram Content Analyzer – це інструмент для автоматизованого збору повідомлень із Telegram-каналів, їхньої обробки, класифікації за категоріями шкідливого контенту та формування аналітичних звітів.

Програма побудована на основі асинхронної взаємодії з Telegram API через Telethon та використовує zero-shot трансформерну модель для класифікації повідомлень без попереднього навчання.

Програмне забезпечення надає такі можливості:

- Підключення до Telegram через API та авторизація користувача.
- Збір повідомлень з вибраних каналів.
- Попередня обробка текстів (очищення від HTML-тегів, посилань, символів).
- Класифікація повідомлень за допомогою багатомовної моделі MoritzLaurer/mDeBERTa-v3-base-mnli-xnli.
- Кешування результатів класифікації для пришвидшення подальшої роботи.
- Генерація звітів у вигляді JSON-файлів і графіків.

Повний вихідний код проєкту доступний у відкритому репозиторії на GitHub: https://github.com/Kyrylo-Kryzhanovskyi/Project .

Б.2 Використані технології

- Python 3.12.9
- Telethon – асинхронний клієнт для Telegram API.
- Transformers (Hugging Face) – робота з трансформерною моделлю.
- Matplotlib – побудова графіків.
- tqdm – відображення прогресу обробки даних.

- asyncio – асинхронна взаємодія із сервером Telegram.
- json, os, re – обробка даних і файлових структур.

Б.3 Структура проєкту

```
/project-root/
├── main.py                # Основний файл запуску
├── config.py              # Конфігурація з API ID та Hash
├── telegram_module.py     # Взаємодія з Telegram
├── preprocessing.py       # Очищення текстів
├── classifier.py          # Класифікація повідомлень
├── report_generator.py    # Формування звітів і графіків
├── utils.py               # Допоміжні функції
├── /model/                # Збережена модель
├── /data/                 # Тимчасові файли
├── /reports/              # Готові звіти
└── requirements.txt       # Список залежностей
```

Папки /data/ та /reports/ створюються автоматично під час роботи програми.

Б.4 Встановлення

1. Завантаження проєкту:

Клонування або завантаження архіву з GitHub-репозиторію: https://github.com/Kyrylo-Kryzhanovskyi/Project

2. Створення віртуального середовища (рекомендовано):

Щоб уникнути конфліктів між бібліотеками, у каталозі з проєктом рекомендується створити окреме віртуальне середовище:

```
python -m venv
```

Активація середовища:

- На Windows:

  ```
  venv\Scripts\activate
  ```

- На Linux/MacOS:

  ```
  source venv/bin/activate
  ```

3. Встановлення залежностей:

Для коректної роботи програми перед встановленням залежностей рекомендується оновити менеджер пакетів pip до останньої версії:

```
python -m pip install --upgrade pip
```

Після активації середовища встановіть необхідні бібліотеки:

```
pip install -r requirements.txt
```

Також необхідно додатково встановити бібліотеку torch (PyTorch), оскільки вона не встановляється автоматично через requirements.txt через особливості збірки для різних систем (CPU або GPU, різні версії CUDA).

Тому її потрібно встановити вручну:

- Для процесорної (CPU) версії:
  ```
  pip install torch torchvision torchaudio
  ```
- Для роботи з відеокартою Nvidia (GPU, CUDA):
  Оберіть команду відповідно до вашої системи на офіційному сайті https://pytorch.org/get-started/locally/. Наприклад, для CUDA 11.8:
  ```
  pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118
  ```

Якщо ви не знаєте, чи підтримує ваша відеокарта CUDA, оберіть перший варіант (CPU).

4. Налаштування API Telegram:

Створіть файл config.py або відредагуйте існуючий у корені проєкту зі своїми обліковими даними:

```
API_ID = 'YOUR_API_ID'
API_HASH = 'YOUR_API_HASH'
SESSION_DIR = 'sessions'
```

Щоб отримати API_ID та API_HASH, необхідно зареєструвати застосунок на офіційному порталі Telegram: https://my.telegram.org/ .

Б.5 Використання

Для запуску програми запустіть основний скрипт у командному рядку:

```
python main.py
```

Після запуску програма запропонує пройти авторизацію через Telegram (ввести номер телефону та код підтвердження). Після успішного підключення можна вибрати потрібні канали для аналізу.

Після завершення аналізу результати будуть збережені:

- У папці /data/ – зібрані повідомлення.
- У папці /reports/ – сформовані звіти з класифікацією та графіками.

Б.6 Вимоги

- Python 3.12.9 або новіший. Рекомендовано 3.12.9.
- Активний обліковий запис Telegram.
- Робота у інтерфейсі командого рядка.

ДОДАТОК В

РИСУНКИ, ЯКІ ОПИСУЮТЬ TELEGRAM CONTENT ANALYZER

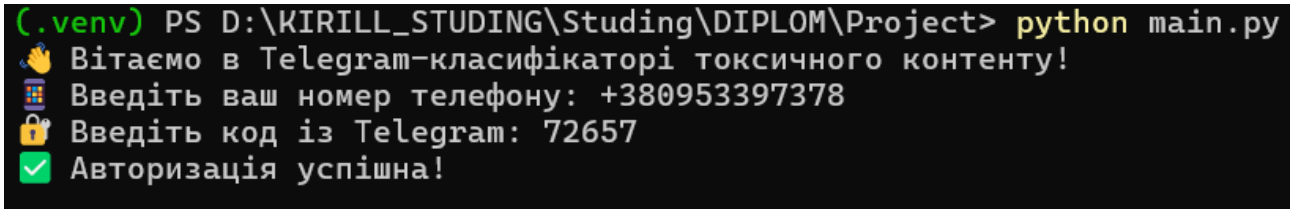

Рисунок В.3 – Приклад авторизації користувача в Telegram через API

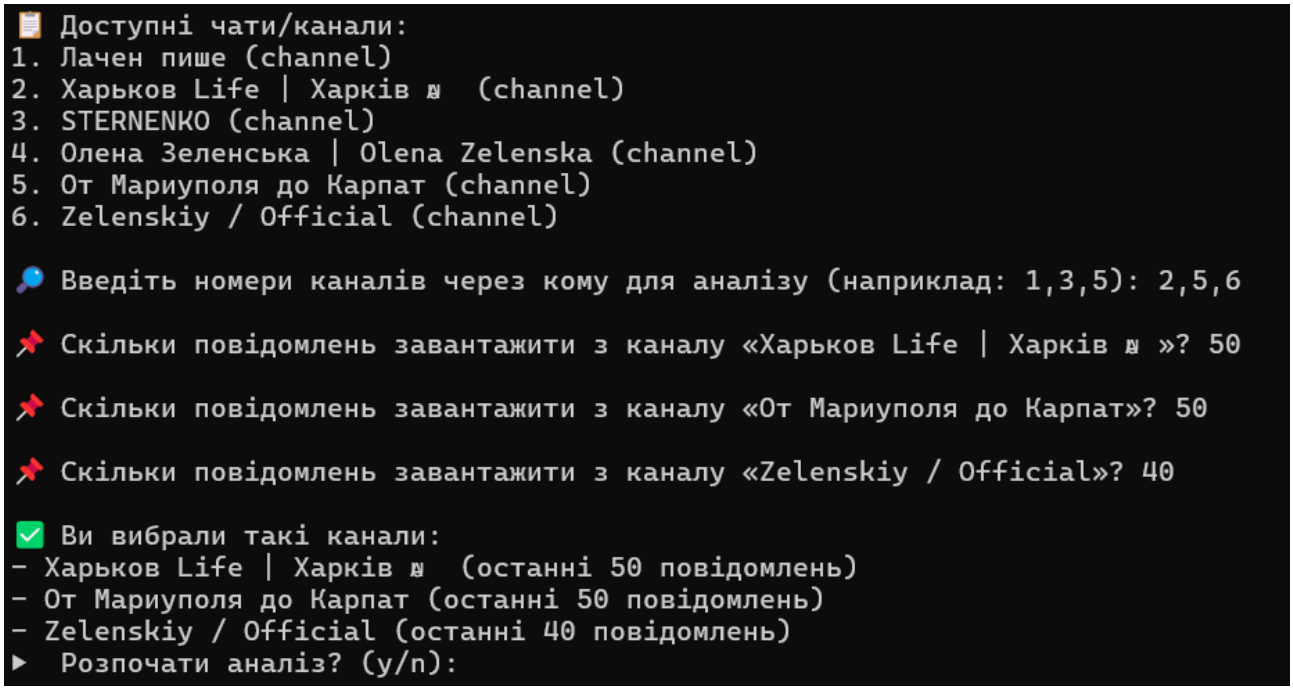

Рисунок В.4 – Приклад вибору каналу та зазначення кількості повідомлень для аналізу

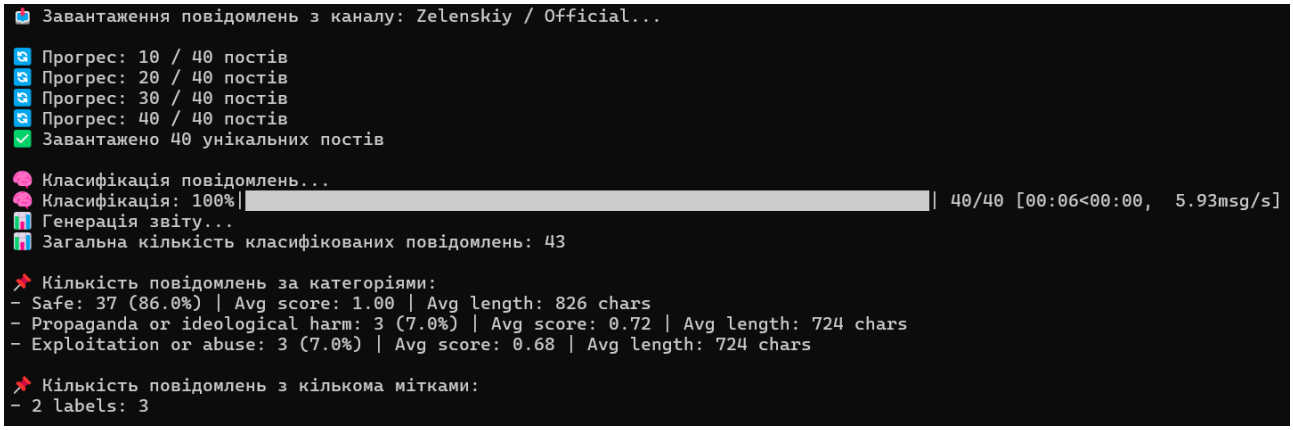

Рисунок В.5 – Приклад процесу класифікації повідомлень та формування короткого підсумкового звіту

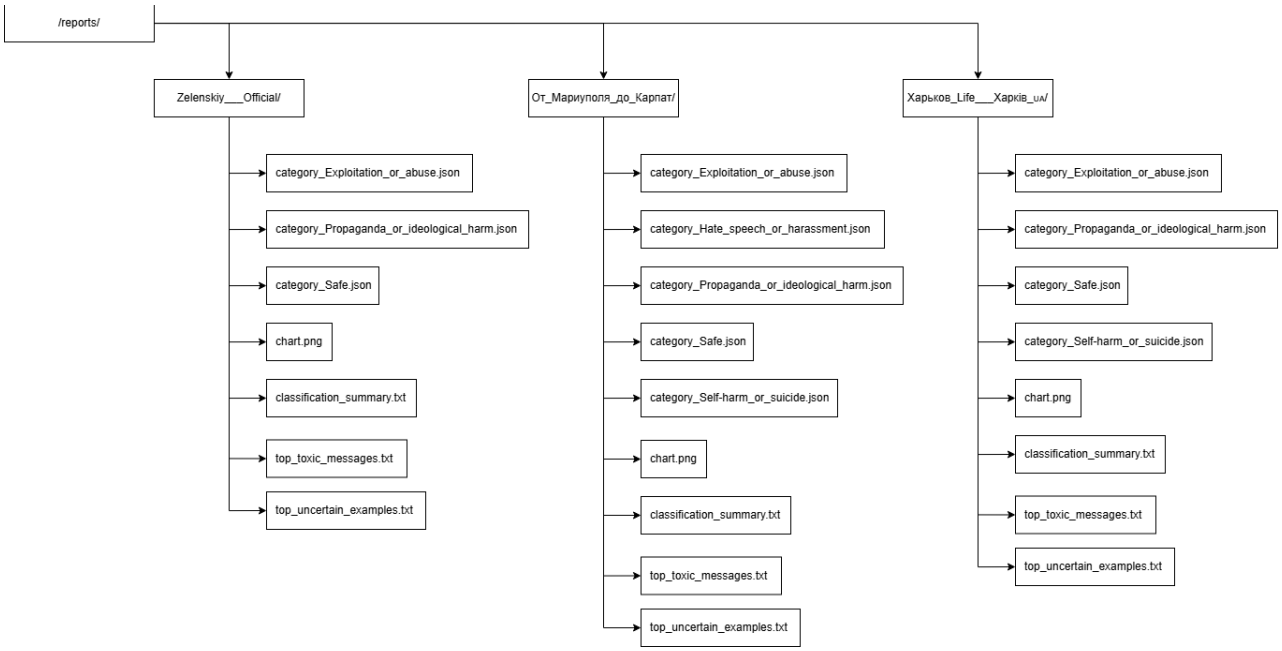

Рисунок В.6 – Приклад структури зберігання згенерованих звітів у директорії /reports/

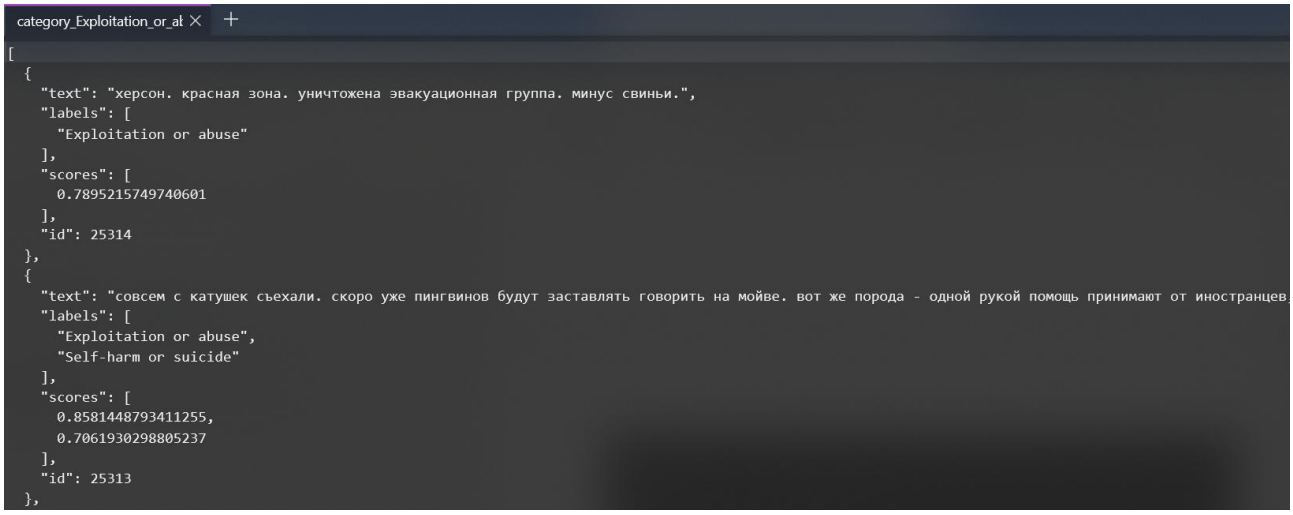

Рисунок В.7 – Приклад вмісту JSON-файлу звіту за категорією Exploitation_or_abuse (category_Exploitation_or_abuse.json)

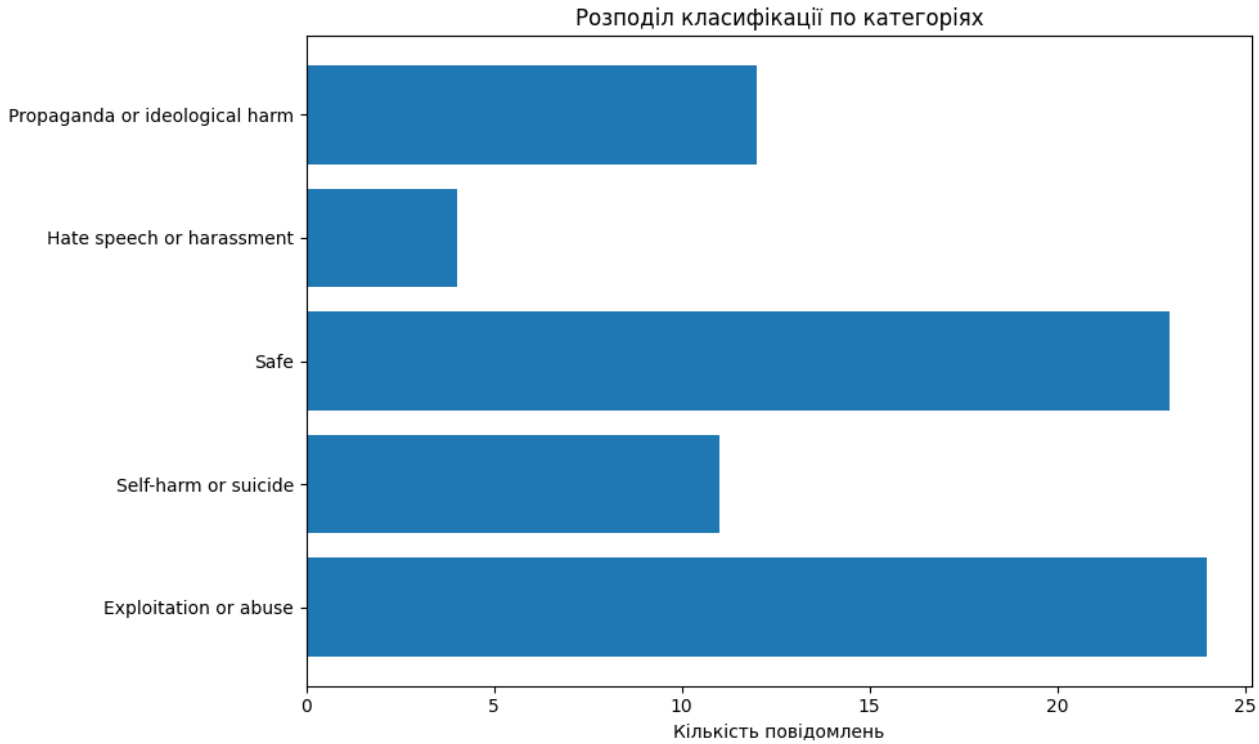

Рисунок В.8 – Приклад діаграми розподілу повідомлень за категоріями шкідливого контенту (chart.png)