Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА

на тему *Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей*

Рівень вищої освіти *другий (магістерський)*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент 2 курсу, група КНм-24-1* _____ *Владислав АНДРОЩУК*
Курс, група виконавця — Підпис — Ім'я, ПРІЗВИЩЕ

Керівник: *Ph.D., ст. викл. кафедри КН* _____ *Марина МОЛЧАНОВА*
Науковий ступінь, посада — Підпис — Ім'я, ПРІЗВИЩЕ

Нормоконтроль: *к.т.н., доцент кафедри КН* _____ *Руслан БАГРІЙ*
Науковий ступінь, посада — Підпис — Ім'я, ПРІЗВИЩЕ

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор _____ *Олександр БАРМАК*
Підпис — Ім'я, ПРІЗВИЩЕ

16 грудня 2025 р.

Хмельницький 2025

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *магістр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*
_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*28*» *серпня* *2025* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

1. Тема кваліфікаційної роботи магістра: *«Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей».*

2. Завдання видано студенту *Владиславу АНДРОЩУКУ*
(Ім'я, ПРІЗВИЩЕ)

3. Керівник роботи *старший викладач кафедри КН Марина МОЛЧАНОВА*
(Ім'я, ПРІЗВИЩЕ)

4. Затверджені наказом університету від «*25*» *серпня* *2025* р. № *65.*

5. Дата видачі завдання студенту: «*28*» *серпня* *2025* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Метою кваліфікаційної роботи є підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб'єктів впливу та спрямованості взаємодії на основі трансформерних моделей. Для досягнення мети слід вирішити такі задачі: дослідити сучасний стан області виявлення кібербулінгу;виконати аналіз наукових досліджень предметної області; розробити метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей; виконати підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу; здійснити програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей; виконати дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

7. Календарний план виконання кваліфікаційної роботи:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження теми кваліфікаційної роботи з керівником, складання календарного графіка виконання роботи*	*вересень 2025*	*Виконано*
2	*Ознайомлення з предметною областю, аналіз існуючих методів і моделей, формулювання мети та завдань дослідження, визначення об'єкта й предмета дослідження*	*вересень 2025*	*Виконано*
3	*Розробка методу чи моделі для вирішення обраного завдання, опис архітектури рішення*	*жовтень 2025*	*Виконано*
4	*Програмна реалізація методу чи моделі*	*жовтень 2025*	*Виконано*
5	*Дослідження ефективності та експериментальна перевірка результатів, порівняння з відомими підходами*	*листопад 2025*	*Виконано*
6	*Написання пояснювальної записки, оформлення відповідно до вимог, врахування зауважень керівника*	*листопад 2025*	*Виконано*
7	*Підготовка презентаційних матеріалів та попередній захист*	*листопад 2025*	*Виконано*
8	*Перевірка пояснювальної записки на відповідність вимогам оформлення (нормоконтроль) та перевірка на академічну доброчесність. Отримання відгуку керівника та рецензії.*	*грудень 2025*	*Виконано*
9	*Публічний захист кваліфікаційної роботи*	*грудень 2025*	*Виконано*

Виконавець: *студент групи КНм-24-1* (Група виконавця) _____ (Підпис) *Владислав АНДРОЩУК* (Ім'я, ПРІЗВИЩЕ)

Керівник: *ст. викладач каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Марина МОЛЧАНОВА* (Ім'я, ПРІЗВИЩЕ)

Реферат

Кваліфікаційна робота присвячена вирішенню науково-технічної задачі автоматизованого виявлення суб’єктів впливу кібербулінгу у цифрових комунікаціях на основі трансформерних моделей. Результатом роботи є метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

Актуальність теми визначається зростанням частоти агресивної взаємодії у цифрових комунікаціях та недостатньою здатністю автоматизованих систем виявляти приховані форми впливу у текстовому середовищі. Кібербулінг все частіше проявляється не через відверто образливу лексику, а через непрямі мовленнєві дії, що формують тиск, підтримують ескалацію або задають тональність дискурсу. Більшість існуючих підходів зосереджені на класифікації повідомлень як токсичних або нейтральних, тоді як структура взаємодії між авторами цих повідомлень залишається поза увагою.

У таких умовах важливим стає аналіз того, як різні учасники текстової комунікації впливають на розвиток агресивного контенту незалежно від їхньої персональної ідентифікації. Застосування трансформерних моделей відкриває можливість урахування контексту, послідовності та функціональної ролі мовленнєвих актів, що дозволяє досліджувати вплив суб’єктів не на рівні особи, а на рівні текстової поведінки. Розроблення методу багаторівневого виявлення таких суб’єктів забезпечує перехід від ізольованого аналізу висловлювань до моделювання комунікативної динаміки, що підвищує ефективність моніторингу, інтерпретації та попередження кібербулінгових проявів.

Мета і задачі роботи. Метою кваліфікаційної роботи є підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб’єктів впливу та спрямованості взаємодії на основі трансформерних моделей.

Для досягнення мети слід вирішити такі задачі:

– дослідити сучасний стан області виявлення кібербулінгу;

– виконати огляд сучасних методів та засобів виявлення кібербулінгу та суб’єктів впливу;

– виконати аналіз наукових досліджень предметної області;

– розробити метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконати підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу;

– здійснити програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконати дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

Об’єкт дослідження. Процес автоматизованого виявлення кібербулінгу та його суб’єктів у текстових комунікаціях.

Предмет дослідження. Моделі, методи та засоби обробки природної мови для виявлення кібербулінгу та його суб’єктів у текстових комунікаціях.

Методи дослідження, що використані для вирішення поставлених завдань є наступними: методи нейромережевого аналізу тексту для виявлення кібербулінгу та його суб’єктів, методи математичної статистики для оцінювання ефективності запропонованого підходу.

Наукова новизна одержаних результатів полягає у розробленні методу багаторівневого виявлення кібербулінгу, який забезпечує визначення не лише факту агресивної комунікації, а й суб’єктів впливу та спрямованості взаємодії. Запропонований підхід поєднує трансформерні моделі з аналізом комунікативних ролей у текстовому середовищі, що дозволяє відтворювати структурну динаміку агресивних повідомлень і підвищувати пояснюваність результатів автоматизованого моніторингу.

Апробація результатів кваліфікаційної роботи та публікації. За темою кваліфікаційної роботи магістра підготовлено до публікації статтю в фаховому виданні категорії Б. Основні наукові й практичні результати роботи доповідались у доповіді «Трансформерне виявлення суб'єктів кібербулінгу за текстовими повідомленнями» на XVII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2025» (м. Хмельницький) 14-15 листопада 2025 року.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4-х розділів, висновків, переліку посилань із 70 найменувань та 8 додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 87 сторінок. У роботі наведено 21 рисунок та 2 таблиці.

Ключові слова: кібербулінг, трансформерні моделі, обробка природної мови, класифікація тексту, інтерпретованість, синтаксичний аналіз, суб'єкти впливу.

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
BERT	Bidirectional Encoder Representations from Transformers
Transformer-LSTM	гібридна архітектура трансформер + LSTM (Long Short-Term Memory)
OLID	Offensive Language Identification Dataset
OffensEval	Offensive Language Evaluation benchmark
SemEval	Semantic Evaluation (міжнародні змагання з обробки природної мови)
AMiCA	багатокласовий корпус/проект для виявлення ролей у соціальних мережах (асистент / агресор / жертва тощо)
GCN	Graph Convolutional Network, графова згорткова нейромережа
DialogueGCN	графова нейромережа для аналізу діалогів
GCN-моделі	моделі на основі графових згорткових мереж
GDPR	General Data Protection Regulation (Загальний регламент про захист даних, ЄС)
DSA	Digital Services Act – нормативний акт ЄС про відповідальність онлайн-платформ
XAI	Explainable AI, пояснюваний штучний інтелект
BiLSTM-CRF	гібридна модель Bidirectional LSTM + Conditional Random Fields для послідовнісної класифікації
ToxicBERT	модифікація BERT для задач детекції токсичних висловлювань

Вступ

Актуальність теми визначається зростанням частоти агресивної взаємодії у цифрових комунікаціях та недостатньою здатністю автоматизованих систем виявляти приховані форми впливу у текстовому середовищі. Кібербулінг все частіше проявляється не через відверто образливу лексику, а через непрямі мовленнєві дії, що формують тиск, підтримують ескалацію або задають тональність дискурсу. Більшість існуючих підходів зосереджені на класифікації повідомлень як токсичних або нейтральних, тоді як структура взаємодії між авторами цих повідомлень залишається поза увагою.

У таких умовах важливим стає аналіз того, як різні учасники текстової комунікації впливають на розвиток агресивного контенту незалежно від їхньої персональної ідентифікації. Застосування трансформерних моделей відкриває можливість урахування контексту, послідовності та функціональної ролі мовленнєвих актів, що дозволяє досліджувати вплив суб’єктів не на рівні особи, а на рівні текстової поведінки. Розроблення методу багаторівневого виявлення таких суб’єктів забезпечує перехід від ізольованого аналізу висловлювань до моделювання комунікативної динаміки, що підвищує ефективність моніторингу, інтерпретації та попередження кібербулінгових проявів.

Мета і задачі роботи. Метою кваліфікаційної роботи є підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб’єктів впливу та спрямованості взаємодії на основі трансформерних моделей.

Для досягнення мети слід вирішити такі задачі:

– дослідити сучасний стан області виявлення кібербулінгу;

– виконати огляд сучасних методів та засобів виявлення кібербулінгу та суб’єктів впливу;

– виконати аналіз наукових досліджень предметної області;

– розробити метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконати підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу;

– здійснити програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконати дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

Об’єкт дослідження. Процес автоматизованого виявлення кібербулінгу та його суб’єктів у текстових комунікаціях.

Предмет дослідження. Моделі, методи та засоби обробки природної мови для виявлення кібербулінгу та його суб’єктів у текстових комунікаціях

Методи дослідження, що використані для вирішення поставлених завдань є наступними: методи нейромережевого аналізу тексту для виявлення кібербулінгу та його суб’єктів, методи математичної статистики для оцінювання ефективності запропонованого підходу.

Наукова новизна одержаних результатів полягає у розробленні методу багаторівневого виявлення кібербулінгу, який забезпечує визначення не лише факту агресивної комунікації, а й суб’єктів впливу та спрямованості взаємодії. Запропонований підхід поєднує трансформерні моделі з аналізом комунікативних ролей у текстовому середовищі, що дозволяє відтворювати структурну динаміку агресивних повідомлень і підвищувати пояснюваність результатів автоматизованого моніторингу.

Апробація результатів кваліфікаційної роботи та публікації. За темою кваліфікаційної роботи підготовлено до публікації статтю в фаховому виданні категорії Б. Основні наукові й практичні результати роботи доповідались у доповіді на XVII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2025» (м. Хмельницький) 14-15 листопада 2025 року.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4-х розділів, висновків, переліку посилань із 70 найменувань та 8 додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 87 сторінок. У роботі наведено 21 рисунок і 2 таблиці.

РОЗДІЛ 1 Дослідження сучасного стану області багаторівневого виявлення суб'єктів впливу кібербулінгу

1.1 Основні поняття та теоретичні засади області комплексного виявлення кібербулінгу

У цифрових комунікаціях кібербулінг дедалі частіше трактується як комплексна форма систематичної агресії, що реалізується із використанням інформаційно-комунікаційних технологій і проявляється через різноманітні типи контенту, зокрема текстові повідомлення, зображення, відео, коментарі, меми та інші медіатексти [1]. На відміну від офлайн-булінгу, цифрове середовище забезпечує сталість і багаторазову відтворюваність агресивних повідомлень, а також їхню потенційно необмежену аудиторію, що істотно підсилює психологічний вплив на жертву. У цьому контексті кібербулінг розглядається не лише як окремі акти образ чи погроз, а як тривалий комунікативний процес, укорінений у специфіці мережевих платформ та їхніх соціотехнічних механізмів.

Міжнародні організації уточнюють базове визначення булінгу як поведінки з чітким наміром завдати шкоди, яка має повторюваний характер і здійснюється за умов асиметрії сил між учасниками взаємодії; ця концептуальна рамка безпосередньо поширюється й на онлайн-взаємодії [2]. У цифровому середовищі дисбаланс сили може формуватися не лише через соціальний статус чи чисельну перевагу, а й через анонімність, алгоритмічне підсилення контенту, швидкість і масштаб його розповсюдження, а також можливість координованих атак з боку груп користувачів. Таким чином, технологічні властивості платформ стають чинниками, що модифікують класичні уявлення про агресію та насильство, ускладнюючи їхнє формальне виявлення й інтерпретацію.

Оновлена інклюзивна дефініція кібербулінгу, запропонована за підтримки UNESCO та представлена на World Anti-Bullying Forum 2023, акцентує увагу на необхідності врахування соціального, культурного та комунікативного контексту, а також множинних драйверів насильницької поведінки в онлайні [3]. Такий підхід підкреслює, що агресивні висловлювання не завжди можуть бути однозначно

ідентифіковані на рівні лексики чи синтаксису, особливо на текстоцентричних платформах, де значну роль відіграють іронія, сарказм, приховані образи та контекстуальні алюзії. Це зумовлює потребу в більш складних аналітичних моделях, здатних інтерпретувати зміст повідомлень у взаємозв'язку з дискурсивним середовищем.

Для ІТ-галузі задача виявлення кібербулінгу виходить далеко за межі простого маркування «токсичності» окремих повідомлень і безпосередньо пов'язана з питаннями безпеки цифрових продуктів, етичного дизайну модераційних систем та управління ризиками для користувачів [4, 5]. Розробка таких систем вимагає поєднання методів обробки природної мови, машинного навчання та аналізу поведінкових патернів із врахуванням вимог до масштабованості, точності й мінімізації хибних спрацьовувань. Водночас помилки в автоматичному виявленні агресивної комунікації можуть призводити як до недозахисту вразливих груп, так і до необґрунтованих обмежень свободи висловлювань.

Регуляторний вимір цієї проблеми посилюється з ухваленням Європейського Digital Services Act, який закріплює обов'язки онлайн-посередників щодо протидії шкідливому контенту, зниження системних ризиків та підвищення прозорості алгоритмічної модерації [6]. У таких умовах автоматизоване, відтворюване та пояснюване виявлення кібербулінгу й агресивної комунікації перетворюється на технічну та нормативну необхідність для розробників і операторів платформ. Це зумовлює актуальність досліджень, спрямованих на формалізацію поняття кібербулінгу та створення інтелектуальних систем, здатних ефективно працювати в реальних умовах цифрових комунікацій.

Кібербулінг може реалізовуватися у різних формах залежно від каналу комунікації, ступеня інтенсивності, цільової спрямованості та способу впливу [7].

Перепалка, або флеймінг, у наукових джерелах трактується як емоційно насичений та конфронтаційний обмін репліками, у якому переважає агресивна лексика, образи або відверто ворожа риторика [8]. Такі взаємодії зазвичай мають імпульсивний, ситуативний характер і виникають у відкритих цифрових середовищах, зокрема чатах, форумах і коментарних стрічках соціальних мереж, де

конфлікт швидко ескалує та стає публічним [9]. Для флеймінгу характерна взаємність агресії, коли обидві сторони залучені до конфлікту, що відрізняє його від більш асиметричних форм онлайн-насильства.

Домагання або систематичні нападки проявляються через повторювані образливі, принизливі чи погрозливі повідомлення, спрямовані на конкретного адресата з метою психологічного тиску [10]. Такі дії можуть реалізовуватися у вигляді серій приватних або публічних текстових повідомлень, коментарів, телефонних дзвінків, спаму чи навмисного перевантаження каналів комунікації [11]. Ключовою ознакою домагань є їхня тривалість і спрямованість на одну особу, що формує стійке відчуття небезпеки та безпорадності у потерпілого.

Наклеп полягає у навмисному поширенні неправдивої, викривленої або принизливої інформації про конкретну особу з метою завдання репутаційної шкоди. У цифровому середовищі такі практики реалізуються через соціальні мережі, блоги, анонімні платформи або групові чати, де інформація швидко тиражується і зберігається тривалий час [12, 13]. Особливу небезпеку становить поєднання наклепу з візуальними матеріалами чи маніпулятивними наративами, що ускладнює спростування неправдивих тверджень.

Самозванство, або імперсонація, відбувається тоді, коли зловмисник отримує несанкціонований доступ до облікових даних жертви або створює фейковий профіль від її імені. Метою таких дій є розміщення провокаційного, компрометуючого чи агресивного контенту, який підриває довіру до особи та завдає їй соціальної або професійної шкоди [14]. У цифрових екосистемах із високим рівнем довіри до ідентичності користувачів імперсонація може мати особливо руйнівні наслідки.

Розголошення особистої інформації охоплює оприлюднення приватних матеріалів без згоди їхнього власника, включно з фотографіями, особистим листуванням, фінансовими, медичними чи сімейними даними. Такі дії часто супроводжуються шантажем, публічним приниженням або системним психологічним тиском [15]. У цифровому середовищі наслідки розголошення посилюються через складність повного видалення інформації та її подальше неконтрольоване поширення.

Ошуканство, або виманювання інформації, реалізується шляхом отримання конфіденційних відомостей під виглядом довірливої чи дружньої комунікації. Після встановлення контакту зловмисник використовує здобуті дані проти потерпілого або передає їх третім особам для подальших зловживань [16]. Такі практики часто поєднують елементи соціальної інженерії та маніпулятивної психологічної взаємодії.

Соціальне відчуження, також відоме як ізоляція або остракізм, стосується навмисного виключення людини з цифрового простору спілкування [17]. Воно може проявлятися через блокування у групових чатах, обмеження доступу до обговорень, формування «чорних списків» або систематичне знецінення участі особи шляхом інформаційного ігнорування [18]. Така форма насильства часто є менш помітною, проте має значний негативний вплив на психоемоційний стан користувача.

Кіберпереслідування, або сталкінг, охоплює нав’язливе та тривале стеження за онлайн-активністю жертви, збір персональних даних і використання їх для погроз, психологічного тиску або підготовки до фізичного переслідування [19]. Ця форма кібернасильства характеризується високим рівнем ризику ескалації з цифрового простору в офлайн-середовище.

Хепіслепінг характеризується поєднанням реального фізичного насильства з його цифровою демонстрацією. Акти агресії фіксуються на відео та поширюються в мережі з метою приниження жертви, отримання соціального схвалення або розваги аудиторії [20]. Цифрове тиражування таких матеріалів багаторазово посилює травматичний ефект насильства.

Онлайн-грумінг полягає у поступовому налагодженні довірливого контакту з неповнолітнім користувачем з метою маніпуляції, примусу до обміну інтимними матеріалами, шантажу або підготовки до зустрічей сексуального характеру [21]. У цифровому середовищі грумінг часто маскується під дружнє або менторське спілкування, що ускладнює його раннє виявлення та запобігання.

Для того, щоб комунікативна поведінка вважалася кібербулінгом, мають бути наявні такі характеристики [22, 23]:

– Систематичність – агресивні дії або повідомлення повторюються або здійснюються з очевидним наміром тиску чи ескалації.

– Наявність учасників – обов’язково фігурує сторони взаємодії: кривдник, жертва та, за потреби, свідки або пасивні спостерігачі.

– Шкідливий вплив – дії або бездіяльність кривдника призводять до приниження, страху, тривоги, емоційної шкоди, порушення соціальних зв’язків або інформаційної ізоляції.

Якщо повідомлення, жарти, зображення або цифрові дії не викликають у адресата негативної реакції, не є принизливими та не мають ознак повторюваності чи цілеспрямованого тиску, вони не розглядаються як кібербулінг [24].

Теоретично кібербулінг дедалі частіше описується не як сукупність ізольованих агресивних висловлювань, а як динамічна взаємодія ролей у межах цифрового дискурсу. У такій взаємодії, поряд із безпосереднім ініціатором агресії та її адресатом, істотне значення мають інші учасники комунікації, здатні як підсилювати, так і стримувати ескалацію конфлікту [25]. Мережевий характер комунікації зумовлює розширення кола залучених осіб, оскільки навіть пасивна присутність аудиторії може впливати на поведінку агресора та сприйняття ситуації жертвою.

У сучасних теоретичних оглядах і емпіричних дослідженнях зазвичай виділяють декілька типових ролей, зокрема спостерігачів-підсилювачів, які схвалюють або ретранслюють агресивний контент, захисників, що намагаються підтримати жертву чи зупинити конфлікт, а також так званих «аутсайдерів», які не беруть активної участі у взаємодії [26]. У цифрових середовищах ці ролі є значно менш стабільними, ніж в офлайн-контекстах, і можуть швидко змінюватися залежно від теми обговорення, соціальних норм конкретної платформи та ситуативних чинників. Користувач, який у одній дискусії виступає нейтральним спостерігачем, в іншій може стати підсилювачем агресії або, навпаки, захисником.

Рольова мінливість у мережевих дискурсах ускладнює формалізацію кібербулінгу та вимагає аналітичних моделей, здатних враховувати не лише зміст окремих повідомлень, а й ширший комунікативний контекст. Саме тому роле-

центричний підхід формує концептуальну основу для багаторівневого виявлення агресивної поведінки в онлайні. На рівні окремого повідомлення фіксується мовленнєвий акт, його емоційне забарвлення та прагматична спрямованість; на рівні діалогу аналізуються адресність, повторюваність та взаємна спрямованість реплік; на рівні взаємодії визначається функціональна роль суб’єкта впливу в динаміці дискурсу та його внесок в ескалацію або деескалацію конфлікту [27]. Така ієрархічна перспектива дозволяє більш точно моделювати механізми кібербулінгу й створює підґрунтя для розробки інтелектуальних систем, здатних виявляти агресію не лише на лексичному, а й на дискурсивному та соціальному рівнях.

Отже, підсумовуючи наведені положення, кібербулінг у цифрових комунікаціях слід розглядати як багатовимірне явище, яке поєднує різні форми агресивних дій, від словесних перепалок і наклепу до ізоляції та грумінгу, та характеризується систематичністю, цілеспрямованістю і наявністю суб’єктної структури взаємодії. Його особливістю є те, що агресія не обмежується окремим повідомленням чи автором, а розгортається у динаміці дискурсу за участі різних ролей: агресора, жертви, підсилювачів, спостерігачів і захисників. Це зумовлює потребу в багаторівневих автоматизованих методах виявлення, здатних не лише фіксувати факт агресивної комунікації, а й інтерпретувати суб’єктно-адресну спрямованість та функціональні ролі учасників, що є необхідною умовою для створення ефективних систем цифрової безпеки та модерації контенту.

1.2 Методи та засоби багаторівневого виявлення суб'єктів впливу кібербулінгу у соціальному медіадискурсі

Сучасні підходи до автоматизованого виявлення кібербулінгу вийшли за межі пост-рівневої «токсичність/не токсичність» класифікації та рухаються до контекстної інтерпретації дискурсу: хто ініціює агресію, на кого вона спрямована, хто її підсилює або гальмує [28]. Цей зсув забезпечили трансформерні моделі, здатні кодувати довгі залежності, прагматичні сигнали й ланцюги взаємодій у розмовах. Огляд 2024-2025 років фіксує стійку перевагу BERT-подібних і гібридних

Transformer-LSTM архітектур у задачах кібербулінгу, включно з крос-датасетною узагальнюваністю; водночас для ресурсно обмежених мов показано ефективність тонкого донавчання трансформерів на доменно адаптованих корпусах. Це робить трансформери базовим інструментом першого рівня – детекції агресивної комунікації у повідомленні [29].

Ключем до переходу від «факту агресії» до суб'єктно-адресної інтерпретації стало введення в публічні бенчмарки явних ознак спрямованості та цілей. Ієрархія OLID/OffensEval розділяє образливість на націлену й не націлену та виносить окремим підзавданням ідентифікацію мішені образи, що прямо підтримує постановку «ким і на кого спрямовано» [30]. Актуальні ітерації OffensEval підтвердили придатність цієї таксономії для кібербулінгу та споріднених явищ [31].

Другий необхідний компонент – локалізація токсичних фрагментів і пояснюваність. Завдання SemEval-2021 Toxic Spans ввело спан-рівневу анотацію, що дає змогу витягувати саме ті частини висловлювань, які зумовлюють токсичність, тоді як HateXplain додатково маркує цільову спільноту та людські «раціоналі» для навчання моделей із підвищеною інтерпретованістю. Для магістерської теми це означає можливість поєднати класифікацію, виділення спанів і таргет-ідентифікацію в єдиному конвеєрі [32].

Багаторівневе виявлення суб'єктів впливу вимагає моделювання не лише повідомлень, а й взаємодій між учасниками. У корпусах і дослідженнях останніх років рольова перспектива, «агресор / жертва / асистент / підсилювач / захисник / спостерігач», переходить у задачі автоматичної класифікації ролей у реальних соціальних стрічках. Показано, що багатокласові моделі на AMiCA-подібних даних, зокрема трансформери з каскадною або багатозадачною організацією, здатні розрізняти ролі навіть за умов дисбалансу класів, що критично для ідентифікації опосередкованих суб'єктів впливу [33].

Щоб розуміти спрямованість і динаміку, потрібні діалогові та графові моделі. Графові нейромережі для розмов (на кшталт DialogueGCN) кодують міжспікерні залежності та структуру чергування реплік, а в задачах прогнозування «зриву» дискусій GCN-моделі з користувацькою динамікою демонструють покращення над

послідовнісними підходами. Такі засоби природно узгоджуються з нашою метою: відтворити структуру впливу на рівні гілок діалогу й спільнотної взаємодії, а не лише мітити окремі тексти [34].

Окремий клас методів адресує «вплив» у часово-каскадному сенсі – хто ініціює, хто «підхоплює» та хто формує хвіст ескалації. Інтенсивні процеси Гокса та їхні нейронні/динамічні варіанти застосовують для моделювання самозбуджуваних подій у соціальних мережах, кількісно оцінюючи внесок користувачів і мережевих станів у поширення токсичного чи дезінформаційного контенту. Інтеграція таких процесів із текстовими представленнями (в тому числі Transformer-ембеддингами) дає інструмент третього рівня, оцінки «силової» конфігурації суб'єктів впливу в часі [35].

У підсумку, сучасна лінія рішень для багаторівневого виявлення суб'єктів впливу кібербулінгу складається з двох основних шарів: трансформерної класифікації кібербулінгу, та моделей, що інтерпретують ролі й спрямованість впливу. Саме така композиція дозволяє відповідати на питання «ким і на кого спрямовано» із відтворюваною точністю та у форматі, придатному для модераційних систем і наукової верифікації.

1.3 Етичні засади автоматизованого виявлення кібербулінгу та його суб'єктів впливу

Автоматизоване виявлення кібербулінгу та суб'єктів його поширення у цифрових середовищах пов'язане з обробкою висловлювань, поведінкових патернів і динаміки взаємодії між користувачами, що неминуче порушує питання етики, приватності, недискримінації та правомірності використання даних. У документах Ради Європи наголошується, що цифровий моніторинг може застосовуватися лише тоді, коли він спрямований на запобігання шкоді та не призводить до надмірного втручання у комунікацію користувачів [36]. Подібну позицію закріплено у Загальному регламенті про захист даних (GDPR), який вимагає мінімізації обробки

персональної інформації та недопущення ідентифікації осіб на підставі мовленнєвих даних [37].

Правове поле Європейського Союзу, зокрема Digital Services Act, визначає відповідальність онлайн-платформ за виявлення шкідливого контенту, але водночас зобов'язує забезпечувати прозорість алгоритмічних рішень, захист свободи вираження та недопущення надмірної цензури [38]. У звіті UNESCO щодо регулювання цифрових платформ підкреслено, що модераційні системи не повинні відтворювати упередженість або пригнічувати групи користувачів, мовні варіанти чи культурні форми сарказму [39]. Це набуває особливого значення у випадку кібербулінгу, де агресія може бути непрямою, контекстуальною або прихованою за маскуванням емпатії.

Етичні ризики автоматизованого аналізу посилюються можливістю алгоритмічної дискримінації. Дослідження показують, що моделі токсичності на базі трансформерів можуть помилково позначати висловлювання маргіналізованих груп як агресивні через відмінності у лексиці, стилі або жаргоні [40]. Це обґрунтовує потребу в контекстно-чутливому навчанні та застосуванні дебіасингових методів.

Окремий виклик пов'язаний із виявленням суб'єктів впливу у медіадискурсі. На відміну від деанонімізації, рольова ідентифікація повинна ґрунтуватися на лінгвістичних характеристиках і структурі комунікативного ланцюга, а не на персональних даних. У «Ethics Guidelines for Trustworthy AI», розроблених Європейською комісією, зазначено, що системи штучного інтелекту не можуть порушувати право на приватність, створювати профайли осіб або допускати повторну ідентифікацію на основі поведінкових шаблонів [41]. Подібні принципи закріплені й у Законі України «Про захист персональних даних», де підкреслюється заборона обробки чутливих даних без згоди та без визначеної мети [42].

Питання прозорості алгоритмічних рішень також набуває ваги. Підходи explainable AI (XAI) дозволяють інтерпретувати, чому система позначає висловлювання або ролі як ознаки кібербулінгу, знижуючи ризик безконтрольного втручання чи репресивної модерації. Дослідники у своїх рекомендаціях щодо

етичного ШІ наголошує на необхідності поєднання автоматизованих рішень із людським наглядом, можливістю апеляції та перевіркою справедливості моделей.

Український нормативний контекст також містить етичні орієнтири для цифрової безпеки. У Законі «Про освіту» введено визначення булінгу як систематичної агресивної поведінки з метою приниження та соціальної ізоляції, але водночас підкреслено, що реагування має ґрунтуватися на доведеності та недопущенні стигматизації [43]. Це накладає обмеження на алгоритмічні системи, які не повинні формувати соціальні ярлики без контекстної інтерпретації. Також Закон «Про інформацію» вимагає дотримання свободи слова та забороняє блокування висловлювань, які не містять ознак насильницької комунікації чи закликів до дій [44].

Отже, етичні засади автоматизованого виявлення кібербулінгу й суб'єктів його впливу формуються на перетині захисту особистої інформації, недопущення дискримінації, забезпечення прозорості алгоритмів і збереження свободи комунікації. Технології штучного інтелекту можуть бути застосовані лише тоді, коли вони не відтворюють упередження, не ідентифікують особу, не підміняють правову оцінку та не порушують автономію користувача. Поєднання нормативних документів ЄС, рекомендацій міжнародних організацій та українського законодавства створює рамки, у межах яких можлива розробка інтерпретованих і правомірних методів багаторівневого виявлення кібербулінгу.

1.4 Аналіз наукових публікацій за напрямком дослідження

Провідні корпусні ініціативи для автоматизованого аналізу агресивної комунікації поступово рухаються від бінарної класифікації до ієрархічних та пояснювальних завдань. У рамках OffensEval на базі таксономії OLID показано, що трансформерні моделі досягають високих результатів не лише для виявлення образливості, а й для категоризації та ідентифікації мішені. У першій ітерації завдання (2019) найкраща система на підзадачі A (детекція образливості) досягла F1 = 0.829, на підзадачі B (таргетована/нетаргетована образа), F1 = 0.755, а на підзадачі

C (ідентифікація мішені: IND/GRP/OTH) досягнула F1 = 0.660; водночас саме BERT-підходи домінували серед топових учасників [45]. Пояснювальний компонент було розвинуто у SemEval-2021 Task 5 «Toxic Spans», де найкраща команда досягла character-F1 = 70.83% для виділення токсичних фрагментів, підтвердивши практичну реалізованість «спан-рівневого» раціоналізування рішень на базі трансформерів [32]. Окремі системи, що брали участь у тій самій задачі, фіксували F1 ≈ 62.23% при використанні BiLSTM-CRF/«ToxicBERT» підходів, що ілюструє діапазон якості методів та складність спан-детекції [46].

Поряд із загальними корпусами, роботи 2023-2024 років демонструють, що донавчання сучасних трансформерів дає конкурентні результати й на мовно/ресурсно-обмежених даних: наприклад, для бенгальської мови досягнуто F1 = 0.87 із використанням Bangla-BERT/Multilingual-BERT [47]. На твіттер-корпусах низка інженерних рішень показує F1 на рівні 0.91 (зауважимо, що це локальні датасети зі специфічними умовами експерименту) [48]. Огляди та порівняльні дослідження підтверджують стабільну перевагу трансформерних sentence-/cross-encoder-архітектур над традиційними моделями, а також важливість урахування сеансового контексту під час детекції [49, 50]. Для пояснюваних рішень широке застосування має датасет HateXplain, який поєднує мітки класу, ціль спічу та «раціоналі», це дозволяє тренувати моделі детекції разом із модулем інтерпретації (мішень/фрагмент), і тим самим підвищувати довіру до систем модерації [51, 52].

Водночас порівнянність числових результатів потребує обережності: показники суттєво залежать від типу платформи, балансу класів і наявності контексту. Дослідження із залученням «раннього виявлення» у різних доменах (Instagram/Vine) підкреслюють, що контекст та історія взаємодій покращують стабільність моделей відносно ізольованого аналізу постів [53]. У цілому тренд останніх років, перехід від пост-рівневої бінарної класифікації до комбінації детекції, локалізації токсичного спану та таргет-ідентифікації, що відповідає ідеї «інтерпретованого моніторингу» для подальшої модерації.

1.5 Постановка задачі

Метою кваліфікаційної роботи є підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб’єктів впливу та спрямованості взаємодії на основі трансформерних моделей.

Для досягнення мети слід вирішити такі задачі:

– дослідити сучасний стан області виявлення кібербулінгу;

– виконати огляд сучасних методів та засобів виявлення кібербулінгу та суб’єктів впливу;

– виконати аналіз наукових досліджень предметної області;

– розробити метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконати підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу;

– здійснити програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконати дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

РОЗДІЛ 2 Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей

2.1 Підхід до багаторівневого виявлення суб'єктів впливу кібербулінгу

Запропонований підхід до багаторівневого виявлення суб’єктів впливу кібербулінгу ґрунтується на послідовному опрацюванні текстових даних із поєднанням трансформерних моделей і синтаксико-семантичного аналізу (рисунок 2.1). На вхід надходять неструктуровані повідомлення, які розглядаються як потенційні носії агресивної комунікації. Початковий етап полягає у виявленні наявності кібербулінгу за допомогою моделі, здатної класифікувати висловлювання за ознаками образливості, цькування або вербального тиску. Це дозволяє виокремити лише ті фрагменти, що потребують подальшої інтерпретації.

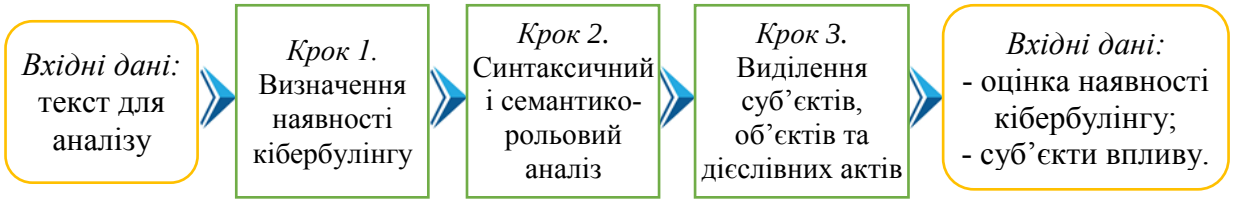


Рисунок 2.1 – Схема та кроки підходу до багаторівневого виявлення суб'єктів впливу кібербулінгу

Після первинної фіксації кібербулінгу текст переходить до рівня синтаксично-семантичного розбору. Застосування парсингу залежностей і моделей рольового аналізу забезпечує реконструкцію граматичної структури висловлювань, визначення предикатів, ідентифікацію лексичних носіїв дії та виявлення семантичного зв’язку між учасниками комунікації. Це створює основу для відокремлення ключових актантів висловлювання.

Кінцева стадія передбачає інтерпретацію взаємодії між учасниками на рівні дії: встановлюється, хто ініціює мовленнєвий акт, на кого він спрямований і яким чином формулюється вплив. Ідентифікація суб’єктів, об’єктів і дієслівних актів дозволяє відтворити комунікативну структуру кібербулінгу та визначити суб’єкти

впливу, що забезпечує можливість подальшого аналізу адресності, динаміки та інтервенційного потенціалу. Результатом такого підходу є структуроване представлення випадку кібербулінгу, що поєднує факт агресії та чітко окреслену рольову конфігурацію.

2.2 Схематичне та формальне представлення методу багаторівневого виявлення суб'єктів впливу кібербулінгу

Метод методу багаторівневого виявлення суб'єктів впливу кібербулінгу призначений для автоматизованої ідентифікації як факту кібербулінгу в комунікації, так і учасників, залучених у відповідну мовленнєву взаємодію (рисунок 2.2). Його застосування дозволяє не лише фіксувати наявність висловлювань з ознаками кібербулінгу, а й встановлювати, хто формує мовленнєвий вплив, на кого він спрямований і в якому контексті реалізується. У межах такого підходу вхідний текст обробляється шляхом послідовної класифікації на предмет кібербулінгу та синтаксико-семантичного аналізу для реконструкції рольових зв’язків між учасниками дискурсу.

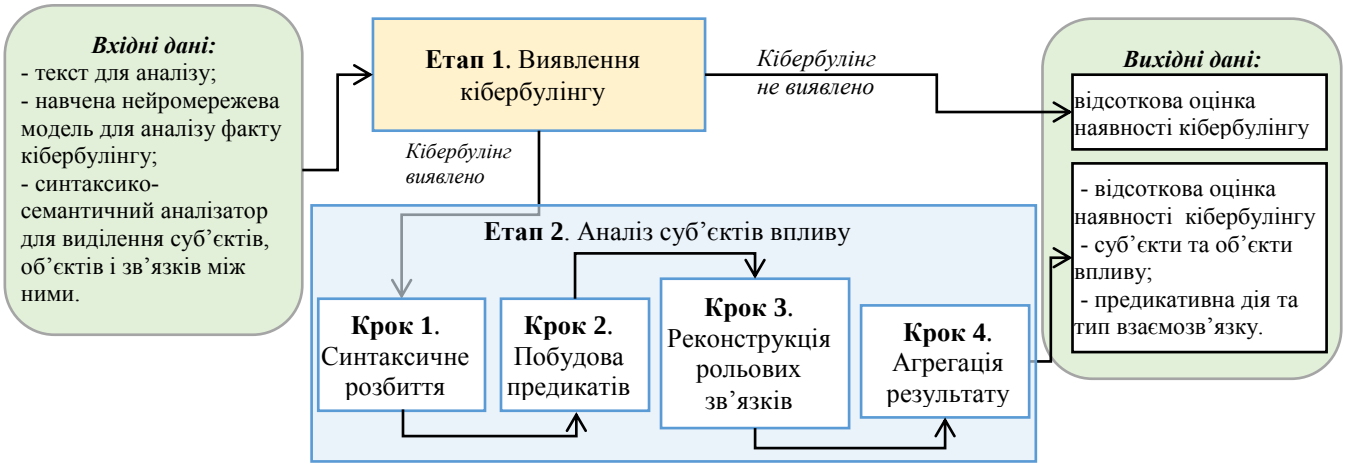


Рисунок 2.2 – Схема методу багаторівневого виявлення суб'єктів впливу кібербулінгу

На вхід надходить текстове повідомлення, яке подається у вигляді послідовності токенів:

$$T=\{w_1,w_2,\ldots,w_n\}, \tag{2.1}$$

де w_i – i-й токен в текстовому повідомленні T. Метою є побудова відображення:

$$M(T)\rightarrow\langle y,R\rangle, \tag{2.2}$$

де $y\in\{0,1\}$ – індикатор наявності кібербулінгу, R – множина семантичних трійок виду (суб’єкт, дія, об’єкт).

На першому етапі здійснюється автоматизоване визначення наявності кібербулінгу за допомогою трансформерної моделі. Вона обчислює ймовірність того, що повідомлення містить ознаки кібербулінгу. Якщо отримане значення перевищує заданий поріг, повідомлення вважається таким, що містить кібербулінг:

Для тексту T трансформерна модель формує значення ймовірності:

$$p=f(T), \tag{2.3}$$

де $f(\cdot)$ – попередньо натренований класифікатор. Рішення приймається за правилом:

$$y=\begin{cases}1, \text{якщо } p\geq\tau\\0, \text{якщо } p<\tau\end{cases} \tag{2.4}$$

де τ – порогове значення.

На етапі 2 спершу здійснюється крок 1, що відповідає за розбиття тексту на множину речень:

$$S=\{s_1, s_2,\ldots,s_k\} \tag{2.5}$$

Для кожного речення виконується розпізнавання залежностей:

$$D(s_i)=\{(head, dependent, relation)\} \tag{2.6}$$

де $\{(head, dependent, relation)$ – множина пар «голова-залежний-тип зв’язку» для речення s_i.

На кроці 2 на основі синтаксичних залежностей відбувається побудова предикатів:

$$Pr=\{pr_1, pr_2,\ldots, pr_m\}, \tag{2.7}$$

де предикатами вважаються дієслова або іменні/прикметникові форми з копулою. Предикативні структури з копулою трактуються як окремий тип предикатів, де лексичним центром є іменник/прикметник, а не службове дієслово-зв'язка.

Реконструкція рольових зв'язків для кожного предиката pr_j визначаються учасники взаємодії:

$$r_j=(subject_j,verb_j,object_j) \tag{2.8}$$

де $subject_j$ – виконавець дії у j-му предикаті, $verb_j$ – дієслівний або предикативний центр взаємодії (сама дія або характеристика), $object_j$ – ціль, на яку спрямована дія або висловлювання (над ким/чим діють).

На кроці 4 відбувається формування результату:

$$R = \{r_1,r_2,\ldots,r_m\} \tag{2.9}$$

Таким чином, метод генерує пару $M(T)=\langle y,\ R\rangle$, де y забезпечує бінарну оцінку наявності кібербулінгу, а R – його рольова інтерпретація у вигляді суб'єктно-об'єктних зв'язків.

Це забезпечує виявлення суб'єктів, об'єктів та природи мовленнєвого акту, що уможливлює подальше моделювання впливу, моніторинг комунікацій і формування інтерпретованих результатів для систем модерації, аналітики або дослідження онлайн-взаємодій.

2.3 Опис й структура навчальної вибірки

В межах дослідження використано 2 набори даних – Cyberbullying Classification [54]. Цей датасет складається виключно з англомовних твітів і містить понад 47 тисяч текстів, кожен із яких позначений однією з шести категорій кібербулінгу: за ознакою віку, етнічної належності, гендерної приналежності, релігії, іншими проявами агресії, а також класом «не кібербулінг». Кожна група містить приблизно однакову кількість прикладів (близько 8 тисяч), що забезпечує збалансованість вибірки. Особливістю корпусу є природна форма мовлення: короткі повідомлення, сленг, орфографічні варіації, сарказм і розмовна лексика, характерні для Twitter. Завдяки цьому датасет добре підходить для навчання моделей як на

виявлення кібербулінгу, так і на визначення його спрямування щодо певних соціальних груп. Водночас у ньому відсутня структурована інформація про суб’єктів і об’єктів впливу, що обмежує можливість аналізу рольових зв’язків.

Ще одним датасетом є «Cyberbullying Detection» [55]. Цей датасет об'єднує тексти з різних платформ, що дозволяє враховувати різні стилі комунікації (коментарі, обговорення, повідомлення, твіти). Він використовується переважно для двокласової класифікації: кібербулінг / не кібербулінг. До корпусу включено висловлювання, що містять агресію, ненависть, образи, токсичність або ворожу лексику. Оскільки дані походять з різних джерел, спостерігається мовна і тематична варіативність, що може покращити узагальнювальну здатність моделі. Недоліком є наявність дисбалансу між класами та відсутність деталізації типу агресії або її адресності. Такий корпус доцільно використовувати як базовий шар для загального виявлення токсичності або як початковий етап донавчання нейромережевої моделі.

2.4 Критерії оцінки ефективності нейромережевої моделі виявлення кібербулінгу

Оцінювання ефективності моделі виявлення кібербулінгу спирається на комплекс метрик, що відображають точність класифікації, здатність виявляти релевантні випадки та стабільність результатів у різних умовах. Precision визначає, яка частка повідомлень, позначених моделлю як агресивні або образливі, справді належить до ворожого контенту. Ця характеристика вказує, наскільки добре система уникає хибних спрацьовувань (false positives), що є критично важливим у контексті автоматичної модерації, де неправомірне блокування чи маркування повідомлень може викликати етичні та комунікаційні проблеми [56].

Recall показує, яку частку реальних проявів кібербулінгу модель здатна виявити. Цей показник характеризує схильність уникати пропусків (false negatives) – ситуацій, коли образливе висловлювання лишається непоміченим. Для завдань онлайн-безпеки Recall часто вважається вирішальним, оскільки невиявлений агресивний контент може мати серйозні психологічні або соціальні наслідки [57].

F1-score використовується як збалансована міра якості, що одночасно враховує Precision і Recall. Це особливо доречно в тих випадках, коли позитивний клас (bullying/offensive) є менш представленим або потребує більшої уваги. У багатокласових сценаріях може застосовуватися macro-, micro- або weighted-усереднення, залежно від цілей аналізу та співвідношення класів [58].

Accuracy відображає частку правильно класифікованих прикладів загалом, але у випадках дисбалансу даних її значення може вводити в оману, оскільки висока точність можлива навіть тоді, коли модель майже ігнорує рідкісні, проте значущі класи. Тому цей показник не розглядається як самодостатній і використовується разом з іншими метриками [59].

Confusion matrix дозволяє деталізовано проаналізувати типи помилок: які саме випадки модель неправильно розпізнає як нейтральні або агресивні. Такий підхід сприяє діагностиці слабких місць, оцінці впливу конкретних класів і налаштуванню порогів прийняття рішень [60].

За наявності ймовірнісного виходу моделі враховується AUC-ROC, що відображає здатність системи відокремлювати агресивні висловлювання від нейтральних за змінного порогового значення. Це забезпечує розуміння поведінки класифікатора не лише в точці прийняття рішення, а й у ширшому діапазоні можливих сценаріїв [61].

2.5 Стійкість методу до доменних зсувів та обмеження застосовності

Запропонований підхід орієнтований на аналіз англомовних повідомлень короткого формату та поєднує нейромережеву класифікацію з подальшим синтаксичним аналізом для виділення відношень типу «підмет-дієслово-об’єкт/комплемент». У реальних умовах експлуатації вхідні дані формуються різними платформами комунікації, які відрізняються довжиною повідомлень, частотою неформальної лексики, наявністю емодзі, хештегів, скорочень, а також змішаним використанням регістрів, орфографічних відхилень і стилістичних конструкцій. Такі відмінності породжують доменні зсуви, за яких статистичні

властивості текстів у цільовому середовищі можуть істотно відрізнятися від розподілу, на якому було навчено модель. Унаслідок цього можливе зниження узагальнювальної здатності класифікатора та погіршення стабільності синтаксичного розбору, що прямо впливає на точність визначення факту кібербулінгу і коректність встановлення суб'єктів впливу.

Стійкість методу до доменних зсувів забезпечується, по-перше, використанням трансформерної архітектури, здатної враховувати контекстні залежності та працювати з варіативними мовними реалізаціями токсичності, включно зі сленгом і непрямими мовними актами. По-друге, застосовується механізм порогового прийняття рішення, який дозволяє гнучко налаштовувати компроміс між хибнопозитивними та хибнонегативними спрацюваннями залежно від умов платформи та цільового сценарію. По-третє, синтаксичний модуль використовується як інтерпретаційний шар, що активується лише за наявності достатніх підстав для токсичного висновку, завдяки чому зменшується накопичення помилок на «чистих» повідомленнях і знижується ризик некоректної атрибуції дійових осіб у нейтральних висловлюваннях.

Водночас межі застосовності методу визначаються природою лінгвістичної неоднозначності та платформозалежними практиками комунікації. Моделі, навчені на даних соціальних мереж, можуть демонструвати нестабільність на довших текстах із багаторівневими дискурсивними структурами або на повідомленнях із високою часткою контексту, що знаходиться поза межами самого висловлювання (попередні репліки, локальні меми, внутрішньогрупові жарти). Для токсичних повідомлень, що реалізуються через іронію, сарказм, завуальовані натяки чи інтертекстуальні посилання, ризик хибної інтерпретації підвищується, оскільки маркери агресії можуть бути непрямими і не мати однозначних лексичних сигналів. Окремим обмеженням є мультимодальність онлайн-комунікації: емодзі, зображення, GIF та інші позатекстові засоби впливу не розглядаються в межах запропонованої постановки, хоча вони здатні змінювати прагматичний зміст повідомлення. Також слід враховувати, що синтаксичний аналіз залежить від якості токенізації та розбору неформального тексту; за наявності масових помилок у пунктуації, відсутності

граматичних зв’язків або нетипового словотвору можливі збої у відновленні відношень підмет-дієслово-об’єкт/комплемент та, відповідно, у виділенні суб’єктів впливу.

З огляду на зазначене, практичне використання методу в різних доменах доцільно супроводжувати перевіркою переносимості на репрезентативних вибірках конкретної платформи та, за потреби, адаптацією моделі шляхом донавчання або налаштування порогів прийняття рішення. Такий підхід дозволяє зберегти узагальнювальну здатність у середовищах з іншими стилістичними нормами, жаргоном і частотами токсичних конструкцій, а також зменшити ризики некоректних висновків у випадках, коли мовні прояви агресії суттєво відрізняються від навчального домену.

2.6 Етичні засади застосування методу виявлення кібербулінгу та суб’єктів впливу

Розроблення та використання інтелектуальних засобів виявлення кібербулінгу потребує дотримання етичних принципів, оскільки результати автоматизованого аналізу можуть впливати на рішення модерації, репутацію користувачів і доступ до цифрових сервісів. Запропонований метод призначений для підтримки аналітика або модератора та не розглядається як самодостатній інструмент встановлення вини чи застосування санкцій. Інтерпретація вихідних даних повинна здійснюватися з урахуванням контексту повідомлення, комунікативної ситуації та можливих прагматичних особливостей, зокрема іронії, сарказму, цитування або репостів, які можуть змінювати семантичну роль автора та адресата.

Ключовим етичним принципом є мінімізація шкоди, що передбачає керування ризиками хибнопозитивних спрацювань, коли нейтральний або захисний вислів може бути помилково інтерпретований як агресивний. З цією метою результати моделі доцільно представляти як імовірнісну оцінку з пояснювальними компонентами, а прийняття практичних рішень реалізовувати за процедурою

«людина в контурі», де остаточний висновок формує відповідальний фахівець. У налаштуваннях системи має бути забезпечено можливість адаптації порога спрацювання під конкретний сценарій використання, щоб уникати надмірної репресивності в середовищах із високою часткою розмовної або емоційно насиченої комунікації.

Не менш важливим є принцип справедливості та недискримінації. Мовні моделі можуть успадковувати з даних навчання соціальні упередження та нерівномірність представленості груп, діалектів або соціолектів, що може призводити до систематично різної якості роботи для різних спільнот. У межах етичного застосування метод потребує перевірки переносимості та якості на різнорідних вибірках, а також періодичного моніторингу помилок із фокусом на потенційно вразливі групи, стилі спілкування й платформи. У випадку виявлення небажаних перекосів необхідно коригувати дані навчання, процедури донавчання або правила використання результатів, не допускаючи автоматичного закріплення дискримінаційних практик.

Принцип приватності та конфіденційності є визначальним для аналізу користувацьких повідомлень. Дані повинні оброблятися в обсязі, необхідному для досягнення мети, з обмеженням строків зберігання та із застосуванням технічних і організаційних заходів захисту. Усі журнали, звіти та експортовані результати мають формуватися таким чином, щоб мінімізувати розкриття персональних даних, а у випадках дослідницького використання забезпечувати знеособлення або псевдонімізацію. За можливості слід уникати збереження сирого тексту, замінюючи його агрегованими метриками або витягнутими ознаками, якщо це не знижує відтворюваності експерименту та не суперечить цілям аналізу.

Окремо слід враховувати відповідальне трактування інтерпретаційних компонентів, зокрема відношень типу «підмет, присудок і об’єкт або комплемент». Такі структури є допоміжним засобом пояснення й не гарантують безпомилкової атрибуції суб’єкта та об’єкта впливу, особливо в умовах неформальної мови, неповних речень або контекстно залежних висловлювань. Тому етично коректне використання інтерпретацій передбачає подання їх як гіпотез, а також наявність

механізмів перевірки й корекції людиною. У застосуванні до модерації доцільно обмежуватися підтримкою пріоритизації перегляду, формуванням пояснювальних звітів і аналітикою тенденцій, не перетворюючи систему на інструмент автоматичного покарання.

Таким чином, етичні засади впровадження методу визначаються поєднанням прозорості, контролю з боку людини, захисту приватності та недопущення дискримінаційних ефектів. Дотримання цих принципів забезпечує коректність і соціальну прийнятність застосування технології, зменшує ризики помилкових рішень та підвищує довіру до результатів автоматизованого аналізу у практичних сценаріях.

Висновки до розділу 2

У розділі обґрунтовано метод багаторівневого виявлення суб’єктів впливу кібербулінгу, що поєднує трансформерні моделі класифікації та синтаксико-семантичний аналіз. Запропонований підхід забезпечує поетапну інтерпретацію текстових повідомлень: від виявлення агресивної комунікації до реконструкції рольових зв’язків між учасниками дискурсу. Така послідовність дає змогу не лише фіксувати факт кібербулінгу, а й визначати його суб’єктну спрямованість і характер мовленнєвого впливу.

Розроблена схема методу формалізує процес обробки тексту як відображення, що повертає бінарну оцінку наявності кібербулінгу та множину семантичних трійок, сформованих на основі предикативної структури речень. Використання dependency-parsing та механізмів відновлення актантних ролей дає змогу автоматизовано виокремлювати учасників взаємодії та встановлювати їх комунікативні функції. Урахування конструкцій із копулою розширює здатність методу інтерпретувати непрямі або приховані форми мовленнєвого впливу.

Застосування спеціалізованих корпусів кібербулінгу дає можливість будувати моделі як для загального виявлення агресивних висловлювань, так і для подальшого донавчання з урахуванням соціальних категорій або типів ворожої

лексики. Аналіз використаних датасетів засвідчує збалансованість одного корпусу за класами та стилістичну різнорідність іншого, що створює умови для комбінованого навчання та адаптації трансформерних моделей.

Показники оцінювання, визначені в межах розділу, формують комплексну систему вимірювання ефективності моделі, орієнтовану як на точність класифікації, так і на здатність виявляти контекстуально значущі прояви агресії. Використання Precision, Recall, F_1-score, Accuracy, матриці помилок та AUC-ROC забезпечує багатовимірне оцінювання та дозволяє адаптувати метод до різних типів дискурсивних даних і рівнів токсичності.

Таким чином, методологічні засади, викладені в цьому розділі, створюють основу для реалізації інструменту, здатного не лише виявляти кібербулінг, а й формувати структуровану модель його акторів та об’єктів, що є ключовим для аналітичних, модераційних і превентивних систем.

Стійкість запропонованого методу до доменних зсувів визначається здатністю трансформерної моделі узагальнювати різні стилістичні реалізації агресії та можливістю адаптації порогового рішення під конкретну платформу, тоді як синтаксичний аналіз доцільно розглядати як інтерпретаційний компонент, чутливий до якості неформального тексту. Водночас межі застосовності зумовлюються контекстозалежністю онлайн-комунікації, поширеністю іронії та імпліцитних форм токсичності, а також мультимодальністю взаємодії, що не враховується в межах поточної постановки, тому практичне впровадження має супроводжуватися перевіркою переносимості на даних цільової платформи та, за потреби, доменною адаптацією моделі.

РОЗДІЛ 3 Проєктування інтелектуальної системи багаторівневого виявлення суб'єктів впливу кібербулінгу

3.1 Вибір засобів розробки інтелектуальної системи

Для реалізації прототипу системи обрано інструменти, які дозволяють поєднати роботу з трансформерними моделями, синтаксичним аналізом тексту та побудовою інтерактивного інтерфейсу. Основою програмної реалізації є мова Python, оскільки вона має зрілий екосистемний набір бібліотек для обробки природної мови, машинного навчання та інтеграції з моделями глибинного аналізу.

Хмарне середовище Google Colab [62] використано як платформу для виконання коду. Воно надає готовий інструментарій для роботи з Python, підтримує імпорт необхідних бібліотек через pip і дозволяє виконувати обчислення без налаштування локального середовища. Таке рішення зручне на етапі дослідження, оскільки не потребує встановлення CUDA, PyTorch чи окремих NLP-фреймворків вручну, а також забезпечує можливість підключення зовнішніх датасетів через Google Drive.

Для класифікації кібербулінгу застосовано бібліотеку Transformers (HuggingFace) [63], яка містить попередньо навчені моделі й інтерфейси для їх використання без додаткової адаптації. Модель для виявлення кібербулінга інтегрована через високорівневий інтерфейс pipeline, що спрощує отримання результатів класифікації.

Для синтаксико-семантичного аналізу тексту використано бібліотеку Stanza, яка забезпечує токенізацію, визначення частин мови, морфологічні ознаки й залежнісне парсування. Завдяки вбудованим моделям, Stanza дозволяє отримувати інформацію про речення, предикати, підмети, об'єкти та інші учасники висловлювання.

Бібліотеки pandas [64] і typing [65] застосовано для структурування проміжних результатів у вигляді таблиць та типізованих структур. Це спрощує обробку даних і передачу результатів між компонентами системи.

Для представлення результатів у вигляді демонстраційного застосунку використано Gradio [66]. Цей інструмент дозволяє будувати веб-інтерфейс без додаткового серверного програмування, що зручно для тестування й візуалізації роботи моделі. Користувач може ввести текст, отримати класифікацію та побачити витягнуті семантичні ролі у табличному форматі.

Поєднання зазначених інструментів забезпечує мінімальні накладні витрати на налаштування й дозволяє зосередитись на логіці методу, а не на конфігурації середовища. Обрані засоби є відтворюваними та доступними, що важливо для міжплатформного використання й подальшого розширення системи.

3.2 Проєктування складових інтелектуальної системи

Схема взаємодії компонентів інтелектуальної системи наведена на рисунку 3.1. У поданій схемі відображено послідовну взаємодію компонентів системи, яка обробляє текст від моменту введення до формування результату. Взаємодія починається з інтерфейсу Gradio UI, через який користувач подає текст. Цей інтерфейс не виконує аналізу, а лише передає дані до центрального модуля – контролера. Контролер координує всі наступні кроки та визначає, які компоненти залучаються залежно від результатів обробки.

Першим етапом аналізу є виклик трансформерного класифікатора, який генерує значення ймовірності наявності кібербулінгу в повідомленні. Значення p оцінюється через блок порогової обробки, де ймовірність зіставляється із порогом τ. Якщо ймовірність є недостатньою, процес обмежується лише формуванням виключного висновку без подальших перетворень, і результат одразу передається на форматування. Якщо ж поріг перевищено, контролер ініціює наступний етап – синтаксико-семантичний аналіз.

Для реалізації структурного розбору використовується Stanza [67], яка забезпечує токенізацію, визначення частин мови та синтаксичних залежностей. Цей модуль спирається на мовні моделі, що можуть доадаптовуватися на основі корпусів, пов’язаних із тематикою агресивної комунікації. Витягнуті залежності

переходять до модуля семантико-синтаксичного структурування, який інтерпретує їх як потенційні підмети, предикати та об’єкти висловлювань. На цьому етапі формується набір трійок, які описують актантні ролі в реченнях.

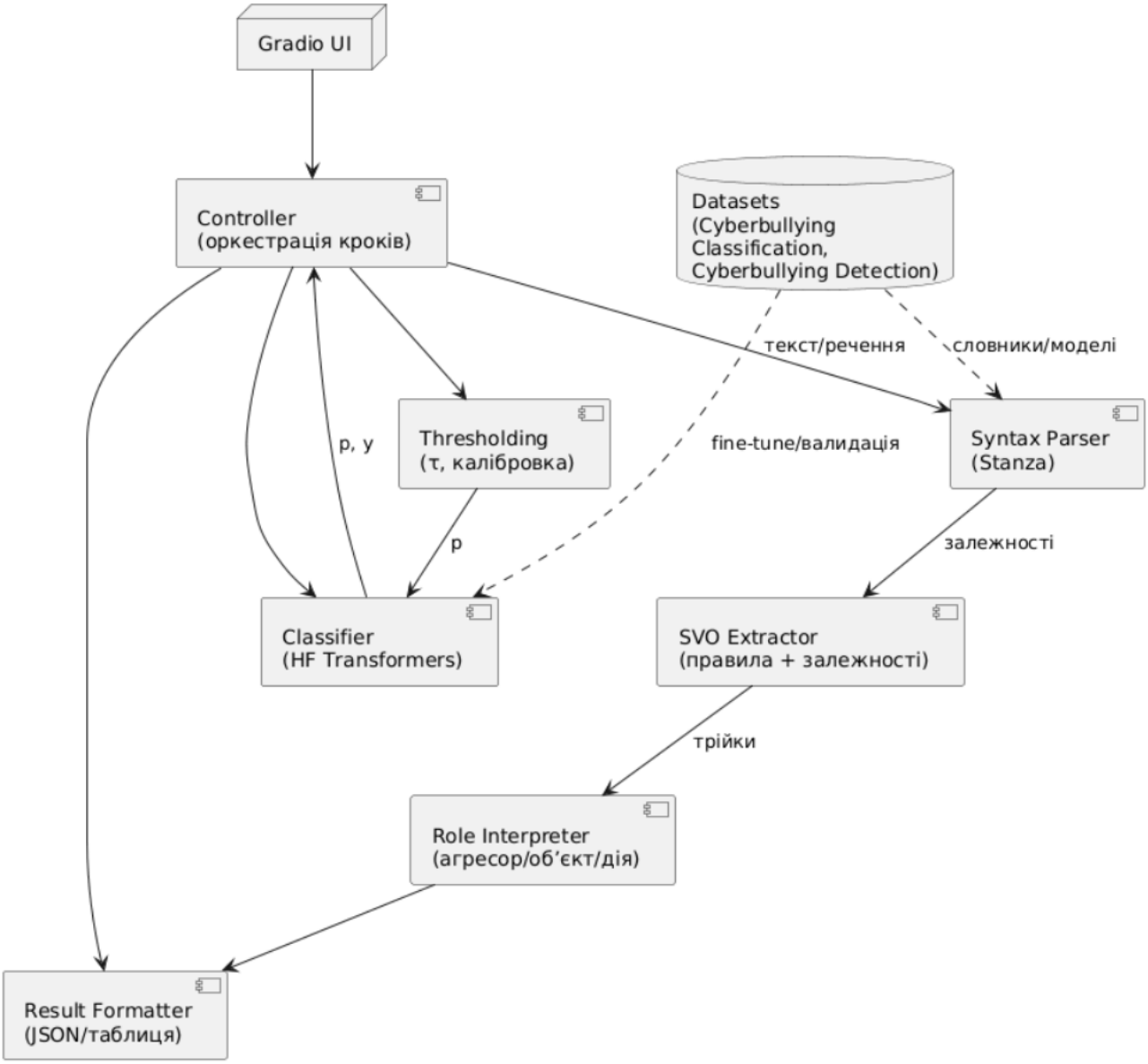


Рисунок 3.1 – Схема взаємодії компонентів інтелектуальної системи

Отримані трійки потрапляють до модуля інтерпретації ролей, де уточнюється, хто виступає мовним ініціатором, проти кого спрямовано висловлювання і яка дія його визначає. Лише після цього контролер передає всю інформацію до модуля форматування. Фінальний результат може містити два рівні

представлення: оцінку наявності кібербулінгу та реконструйовані структури взаємодії, якщо наявність кібербулінгу підтверджено.

Графічні зв’язки демонструють, що контролер утримує управління всіма потоками, а класифікатор і синтаксичний аналізатор працюють не паралельно, а послідовно. Датасети залучені як джерело навчальних прикладів і лексичних моделей, однак не інтегровані в саму логіку виконання під час користувацького запиту. У фінальній точці працює лише генерація результату, що повертається через Gradio UI у вигляді текстового висновку та таблиці ролей.

Схема демонструє, що система поєднує два основні рівні обробки: трансформерну класифікацію та синтаксико-семантичне структурування. Обидва рівні взаємодіють через контролер, який забезпечує послідовність, відбір гілки виконання та узгодженість результатів.

Діаграма варіантів використання (рисунок 3.2) демонструє взаємодію користувача з інтелектуальною системою виявлення суб’єктів впливу кібербулінгу. Користувач ініціює роботу, вводячи текстове повідомлення для аналізу.

Після передання даних система оцінює наявність ознак агресивної комунікації й обчислює ймовірність кібербулінгу. Якщо виявлений рівень відповідає пороговому значенню, запускається більш глибокий шар обробки: формується структура суб’єктно-об’єктних залежностей у форматі Subject - Verb - Object (S-V-O) [68]. Наступним кроком система інтерпретує ролі учасників, встановлюючи, хто ініціює висловлювання, проти кого воно спрямоване і яка дія лежить в основі комунікативного акту. Завершальним етапом є формування результату у вигляді, придатному для перегляду або експорту. Уся взаємодія з боку користувача зосереджується на поданні тексту, отриманні проміжних і фінальних відповідей та, за потреби, збереженні результатів.

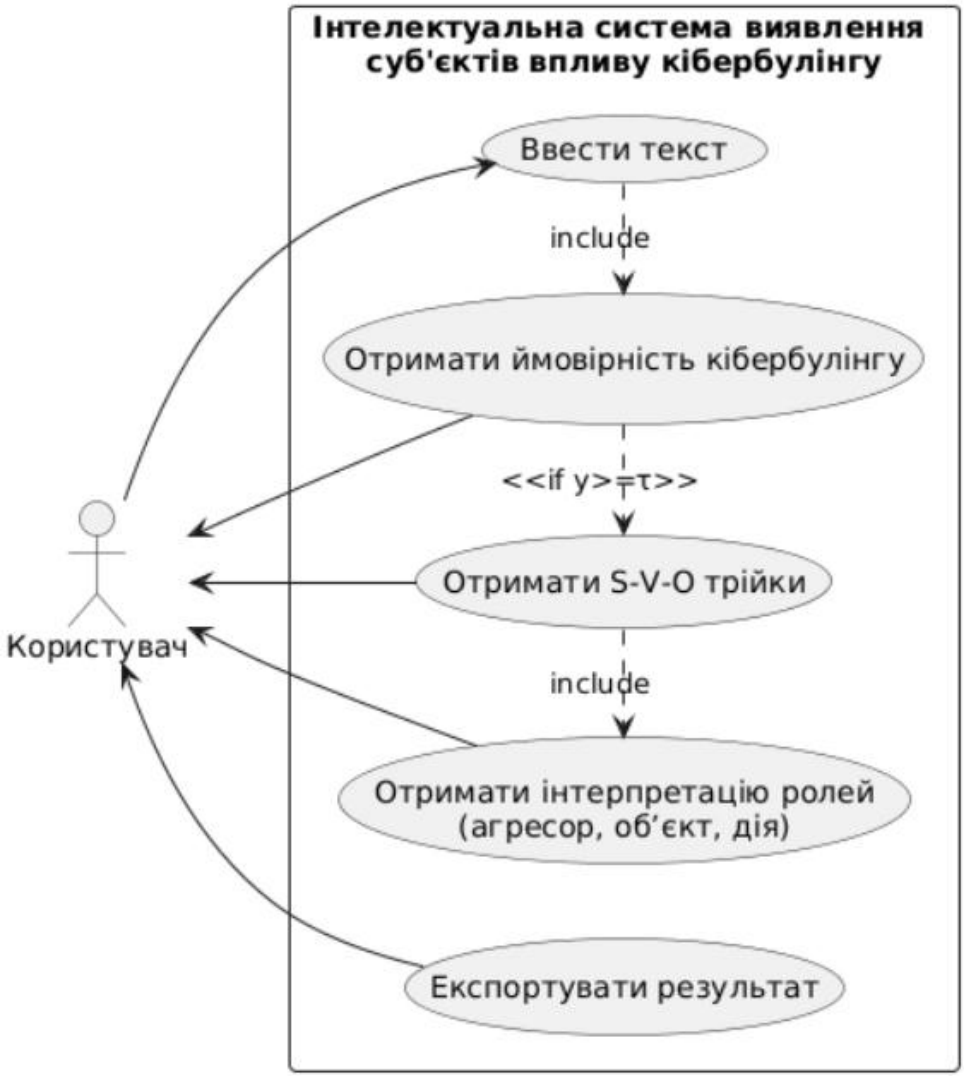


Рисунок 3.2 – Діаграма варіантів використання

Діаграма активності (рисунок 3.3) відображає покроковий рух даних усередині системи – від моменту введення тексту до отримання користувачем інтерпретованого результату. У центрі уваги не лише класифікація, а й умовне розгалуження логіки залежно від того, чи виявлено ознаки кібербулінгу.

Процес починається з надсилання тексту користувачем. Перш ніж здійснювати будь-який аналіз, система виконує нормалізацію та токенізацію, що дозволяє підготувати вхід до класифікатора й синтаксичних моделей. Після цього текст передається до трансформерної моделі, яка повертає оцінку ймовірності p. Саме ця оцінка визначає гілку подальшого виконання.

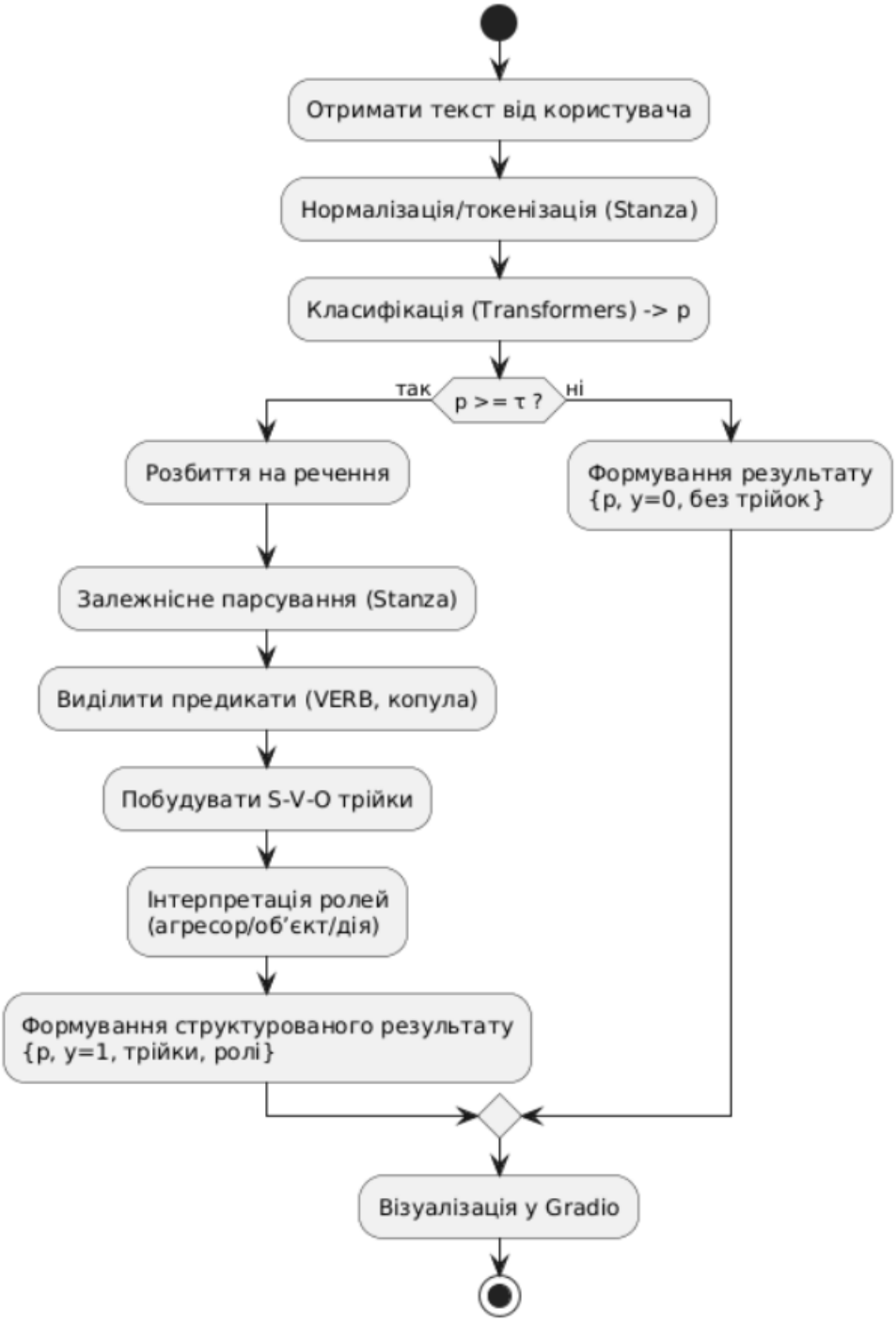


Рисунок 3.3 – Діагарма активності

Якщо значення не досягає порогового рівня τ, система завершує обробку на рівні класифікації і формує відповідь без поглибленого аналізу: y=0, жодні трійки не будуються. Такий сценарій мінімізує зайве використання обчислювальних ресурсів та зменшує кількість хибноінтерпретованих нейтральних повідомлень.

У разі, якщо *p* перевищує поріг, текст переходить до рівня структурного аналізу. Спочатку виконується розбиття на речення, потім – залежнісне парсування,

що виявляє синтаксичні зв’язки між словами. Далі система визначає предикати, включно як із дієсловами, так і з предикативними структурами на основі копули. На цьому фундаменті будуються Subject - Verb - Object представлення, які відбивають базову дію і пов’язаних із нею учасників.

Після побудови таких трійок виконується інтерпретація ролей: розмежовується, хто ініціює дію, на кого вона спрямована та яка подія або характеристика лежить в основі висловлювання. Лише після завершення цих етапів формується структурований результат, де поєднуються показник наявності кібербулінгу, набір трійок і проєкція ролей. Обидва варіанти – як для позитивного, так і негативного рішення – передаються на один і той самий завершальний етап, тобто візуалізацію у Gradio.

У такий спосіб діаграма відображає лінійний потік з умовним розгалуженням: розпізнавання реальних випадків агресії супроводжується повним циклом структурного аналізу, тоді як нейтральний текст не потрапляє в глибшу обробку.

Діаграма послідовностей (рисунок 3.4) відображає динаміку обміну даними між компонентами системи та користувачем під час одного циклу обробки тексту. Уся взаємодія починається з введення повідомлення, яке через інтерфейс передається до контролера. Далі контролер ініціює звернення до трансформерної моделі, яка повертає не лише числову ймовірність, а й бінарне рішення про наявність або відсутність кібербулінгу.

Саме на цьому етапі логіка розгалужується: якщо ймовірність перевищує поріг, контролер переходить до блоку глибшого аналізу. У такому випадку Stanza виконує синтаксичний розбір і формує залежності, на основі яких SVO-екстрактор будує трійки типу “підмет–предикат–об’єкт”. Потім інтерпретатор ролей визначає, хто є агресором, на кого спрямовано висловлювання та якою є дія. Коли інформація сформована, модуль форматування поєднує ймовірність, структурні зв’язки та інтерпретацію ролей у таблицю або JSON-представлення й повертає контролеру, який передає результат користувачу.

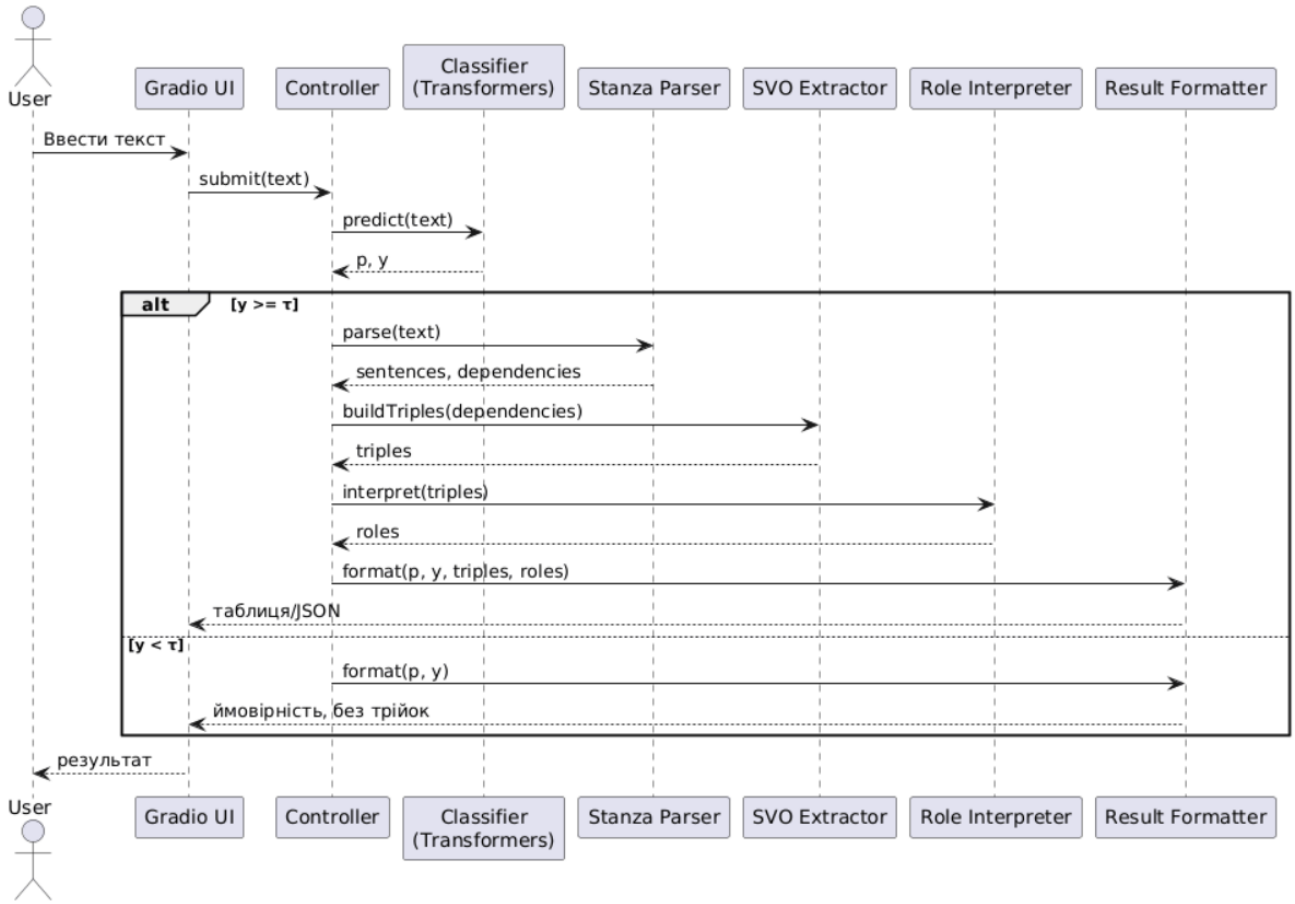


Рисунок 3.4 – Діаграма послідовностей

У випадку, коли порогове значення не досягається, система не виконує синтаксичного аналізу, не будує трійок і не відтворює рольову конфігурацію. Формується лише результат класифікації, який містить оцінку ймовірності без будь-яких додаткових структур. Обидві гілки завершуються поверненням відповіді через Gradio UI до користувача.

Таким чином, діаграма демонструє не просто передачу повідомлень між компонентами, а умовну логіку виконання, за якої система уникає непотрібної глибинної обробки, якщо агресивний контент не виявлено.

3.3 Проєктування інтерфейсу користувача та прототипування екранів системи

Проєктування інтерфейсу користувача виконано з урахуванням прикладного сценарію використання системи в задачах модерації та аналітики: оператор повинен

швидко запустити аналіз, отримати зрозумілий результат класифікації та, за потреби, перейти до пояснювальних компонентів без надлишкового навантаження на увагу. Тому інтерфейс організовано за принципом чіткої структурної ієрархії: фіксована навігація, однозначні назви екранів, логічна послідовність дій і мінімалістична візуальна мова. Окремий акцент зроблено на передбачуваності взаємодії та читабельності: основні елементи керування винесено на видимий рівень, а інформаційні блоки згруповано за призначенням, що зменшує час орієнтації та знижує ризик помилкових операцій під час рутинної роботи.

На рисунку 3.5 подано узагальнений прототип головного екрана системи, який виконує роль стартової точки взаємодії та короткого довідкового огляду.

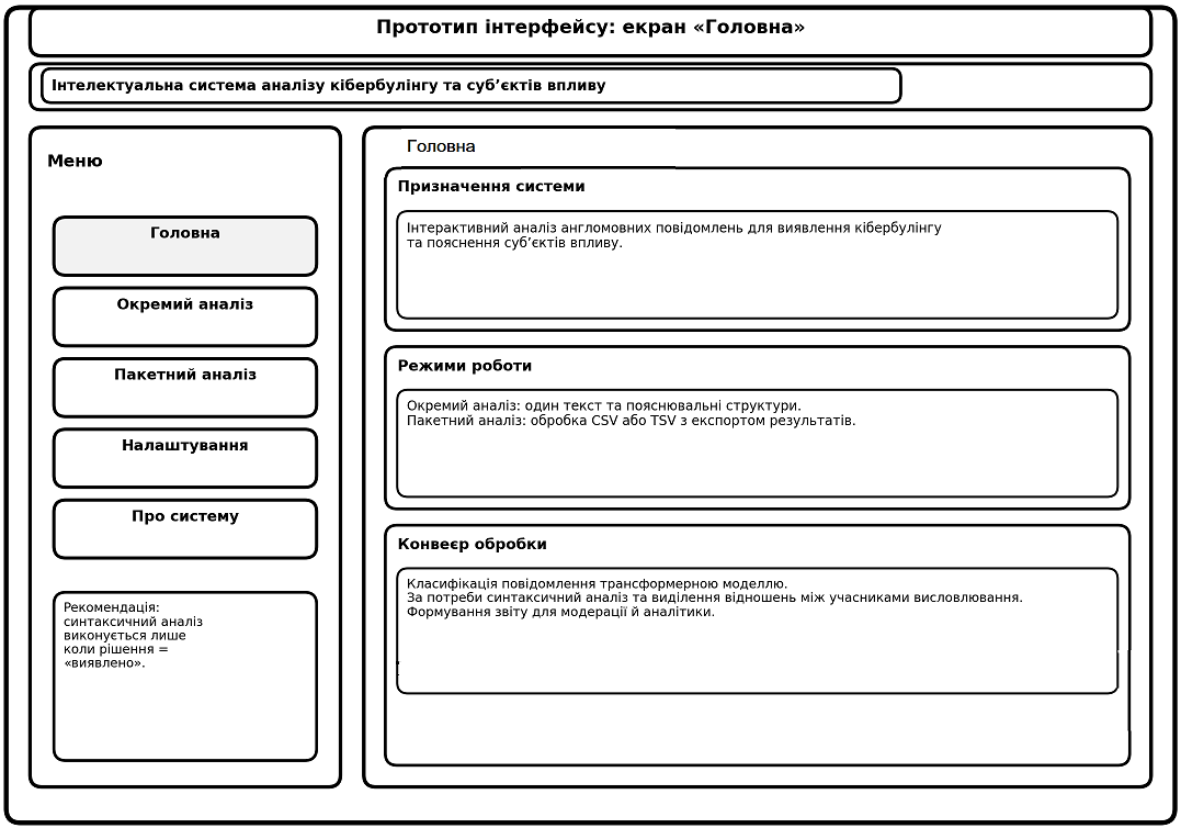


Рисунок 3.5 – Прототип інтерфейсу «Головна»

У верхній частині розміщено заголовок прототипу, нижче наведено шапку з назвою інтелектуальної системи, що забезпечує однозначну ідентифікацію програмного засобу та контекст роботи. Ліва частина макета відведена під вертикальний блок «Меню», який реалізує сталу навігаційну панель і містить

кнопки переходу до основних функціональних екранів: «Головна», «Окремий аналіз», «Пакетний аналіз», «Налаштування» та «Про систему». Таке компонування підтримує однакову модель навігації для всіх режимів і забезпечує швидкий перехід між сценаріями без повернення на попередні сторінки.

У нижній частині лівої панелі розміщено інформаційний блок «Рекомендація», який подає коротку інструктивну підказку щодо виконання синтаксичного аналізу лише у випадку, коли рішення класифікатора свідчить про виявлення кібербулінгу. Цей елемент орієнтований на зменшення зайвих обчислень і зниження ризику некоректних пояснень для нейтральних повідомлень, а також виконує функцію нагадування про логіку роботи інтерпретаційного шару.

Права, основна частина екрана є змістовою панеллю «Головна» і структурована на три послідовні інформаційні секції. Перша секція «Призначення системи» стисло фіксує цільову функцію: інтерактивний аналіз англомовних повідомлень для виявлення кібербулінгу та пояснення суб’єктів впливу. Друга секція «Режими роботи» задає операційну рамку використання, пояснюючи відмінність між режимом аналізу одного тексту з пояснювальними структурами та режимом пакетної обробки таблиць CSV або TSV із подальшим експортом результатів. Третя секція «Конвеєр обробки» описує загальну послідовність процедур: нейромережеву класифікацію повідомлення трансформерною моделлю, умовне виконання синтаксичного аналізу та виділення відношень між учасниками висловлювання, а також формування вихідного звіту, придатного для подальшого використання у модерації або аналітичних задачах. Загальна композиція екрана орієнтована на те, щоб користувач із першого погляду розумів можливості системи, обирав потрібний режим та усвідомлював логіку обробки без необхідності звернення до зовнішньої документації.

На рисунку 3.6 наведено прототип екрана «Окремий аналіз», який реалізує основний інтерактивний сценарій роботи з одиничним повідомленням та формуванням пояснювальних результатів. Загальна композиція екрана зберігає єдину для системи структуру: у верхній частині розміщено службовий заголовок

прототипу, нижче наведено шапку з назвою інтелектуальної системи, а ліва панель використовується як постійний навігаційний блок «Меню».

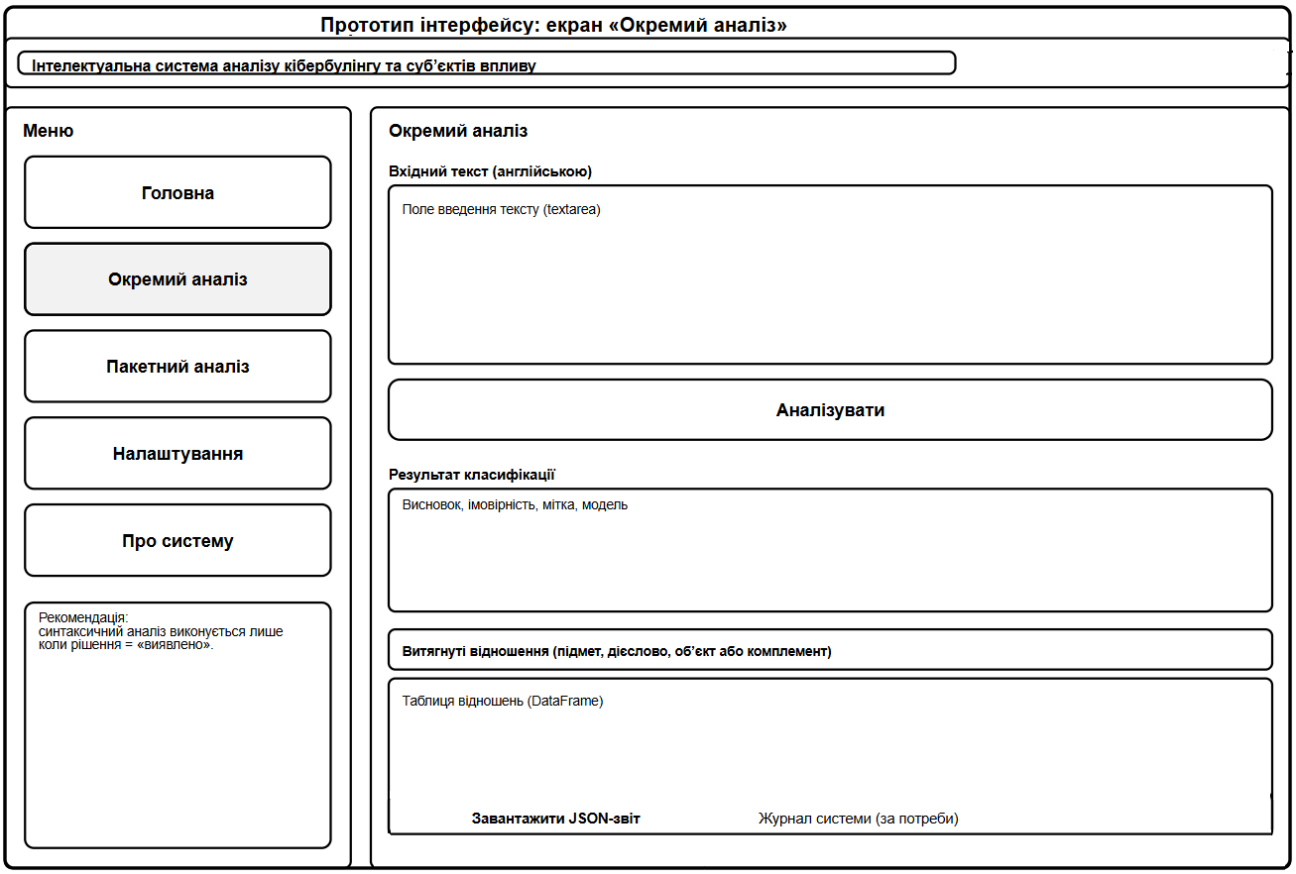


Рисунок 3.6 – Прототип інтерфейсу «Окремий аналіз»

Основна права частина екрана містить робочу область «Окремий аналіз», де елементи керування та результати впорядковано за природною послідовністю виконання. У верхній частині робочої області подано секцію введення «Вхідний текст (англійською)» з великим полем введення, призначеним для вставлення або набору повідомлення, що підлягає аналізу. Безпосередньо під полем введення розміщено центральну кнопку «Аналізувати», яка ініціює обчислювальний конвеєр і є єдиною точкою запуску процесу, що мінімізує ризик помилкових дій і спрощує інструктаж користувача.

Нижче розміщено блок «Результат класифікації», який акумулює підсумкові відомості про оброблене повідомлення. У цьому блоці передбачається відображення формалізованого висновку, імовірнісної оцінки, мітки моделі та посилання на

застосовану конфігурацію, що дозволяє користувачу одночасно бачити результат і контекст його отримання. Далі розташовано пояснювальну секцію «Витягнуті відношення (підмет, дієслово, об’єкт або комплемент)», яка відображає результат синтаксичного аналізу у вигляді таблиці; таблиця інтерпретується як структуроване подання ключових ролей і дій у висловлюванні та використовується для пояснення ймовірного спрямування впливу в повідомленні. У нижній частині робочої області передбачено два службові елементи: кнопку «Завантажити JSON-звіт» для експорту машинозчитуваного результату (зручно для інтеграції в процеси модерації або подальшої аналітики) та блок «Журнал системи (за потреби)», який призначений для фіксації діагностичної інформації у випадках помилок виконання або нестандартних ситуацій.

На рисунку 3.7 подано прототип екрана «Пакетний аналіз», призначеного для обробки множини повідомлень у табличному поданні та формування зведених результатів, придатних для подальшої модерації й аналітики.

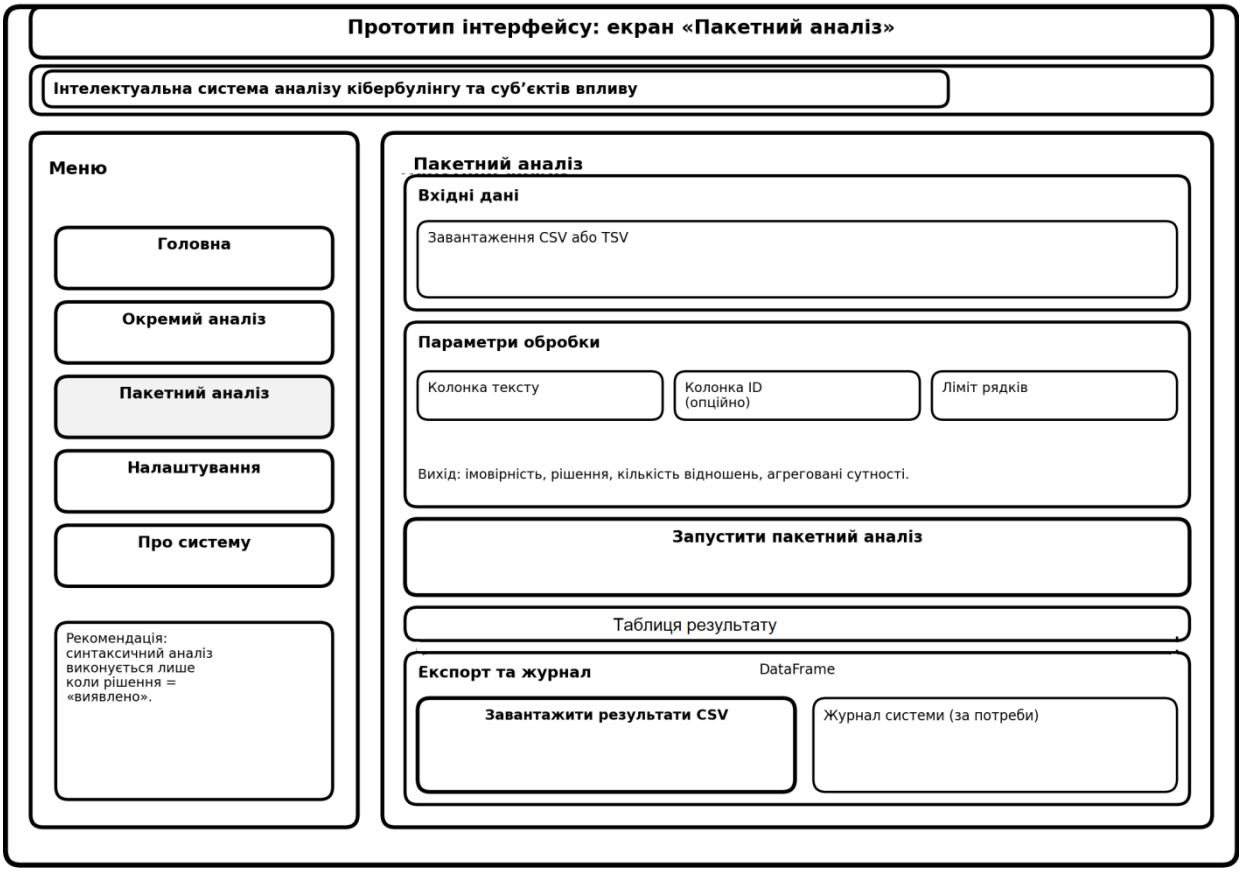


Рисунок 3.7 – Прототип інтерфейсу «Пакетний аналіз»

Компоновка екрана узгоджена з загальною структурою інтерфейсу: зверху розміщено заголовок прототипу та шапку з назвою інтелектуальної системи, ліворуч подано постійну навігаційну панель «Меню» з кнопками переходу між основними екранами, а під меню розташовано інформаційний блок рекомендації щодо умовного виконання синтаксичного аналізу лише у випадках, коли класифікаційне рішення свідчить про наявність кібербулінгу.

У правій робочій області відображено послідовність операцій пакетної обробки. У верхній частині розташовано секцію «Вхідні дані», де передбачено завантаження файлу формату CSV або TSV. Цей елемент задає джерело даних для подальшого аналізу та орієнтований на інтеграцію з типовими експортами з соціальних платформ, журналів модерації або внутрішніх аналітичних вибірок. Нижче подано секцію «Параметри обробки», яка містить поля для налаштування структури вхідної таблиці: вибір колонки з текстом повідомлень, за потреби вказування ідентифікатора запису, а також встановлення ліміту рядків для обробки. Наявність цих параметрів забезпечує переносимість рішення між різними схемами даних і дає можливість контролювати обсяг обчислень у ресурсно обмежених середовищах. У межах цього ж блоку розміщено уточнення щодо складу вихідних даних, зокрема відображення імовірнісних оцінок, підсумкового рішення, кількості витягнутих відношень та агрегованих сутностей, що формує очікування користувача стосовно змісту результатів.

Центральним елементом екрана є кнопка «Запустити пакетний аналіз», яка ініціює обчислювальний конвеєр для всієї завантаженої вибірки відповідно до заданих параметрів. Після виконання обробки результати виводяться в секції «Таблиця результату», представленій у вигляді таблиці, що дозволяє здійснювати швидкий перегляд зведень і проводити первинну верифікацію якості опрацювання. У нижній частині робочої області розміщено секцію «Експорт та журнал», яка поєднує два взаємодоповнювальні компоненти: кнопку «Завантажити результати CSV» для збереження вихідної таблиці у стандартизованому форматі та блок «Журнал системи (за потреби)» для відображення діагностичних повідомлень. Такий підхід забезпечує відтворюваність пакетного експерименту, полегшує

інтеграцію результатів у зовнішні робочі процеси та дозволяє оперативно локалізувати причини збоїв, пов’язаних із некоректним форматом вхідного файлу, відсутністю необхідних колонок або обмеженнями обчислювальних ресурсів.

На рисунку 3.8 представлено прототип екрана «Налаштування», який забезпечує керування конфігурацією нейромережевого модуля, параметрами прийняття рішення та сервісними процедурами реєстрації й діагностики.

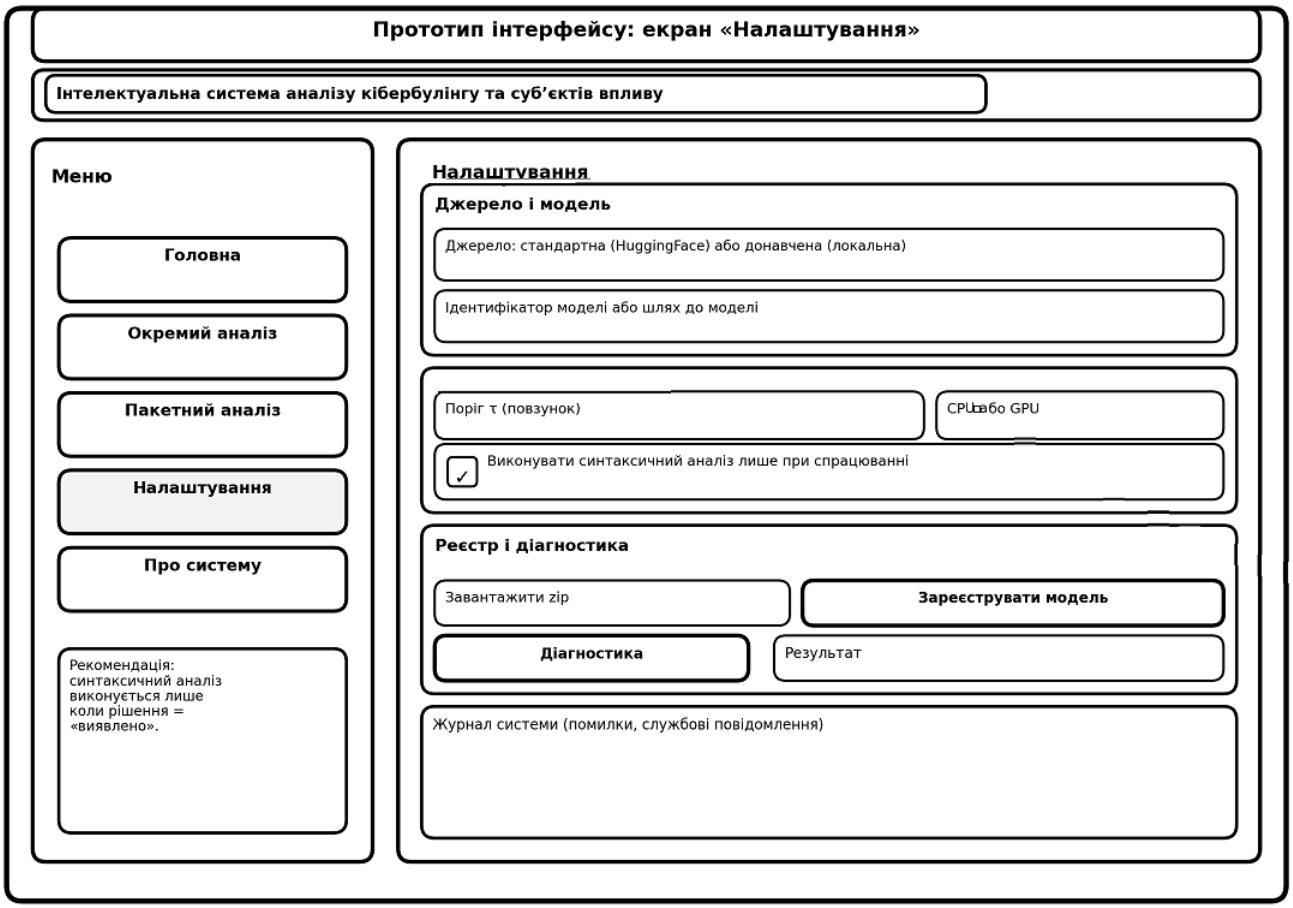


Рисунок 3.8 – Прототип інтерфейсу «Налаштування»

Права робоча область організована як послідовність функціональних секцій, що відповідають задачам налаштування та контролю працездатності. Верхня секція «Джерело і модель» визначає параметри підключення нейромережі: передбачено вибір між стандартною моделлю з репозиторію HuggingFace та донавченою локальною моделлю, що зберігається у файловій системі. Нижче розташовано поле для задання ідентифікатора моделі або шляху до каталогу з моделлю, що забезпечує

гнучкість конфігурації та можливість відтворюваного використання різних варіантів нейромережі в межах однакового інтерфейсу. Такий підхід є принциповим для прикладних сценаріїв, де необхідно порівнювати якість стандартної й адаптованої моделей або оперативно перемикатися між версіями в процесі експлуатації.

Наступний блок налаштувань стосується параметрів прийняття рішення та ресурсного режиму виконання. Окремо виділено керування порогом τ у вигляді повзунка, що дозволяє адаптувати чутливість системи до особливостей домену та вимог конкретного сценарію модерації. Поруч розміщено перемикач обчислювального пристрою «CPU або GPU», який задає режим виконання моделі з урахуванням наявних ресурсів. Нижче подано прапорець «Виконувати синтаксичний аналіз лише при спрацюванні», який реалізує принцип умовної активації інтерпретаційного модуля: синтаксичний розбір і витягування відношень виконуються лише тоді, коли класифікатор приймає рішення про наявність кібербулінгу. Це зменшує обчислювальні витрати та знижує ризик накопичення помилок інтерпретації у випадках нейтральних повідомлень.

Окремою секцією виділено «Реєстр і діагностика», що орієнтована на супровід експлуатації системи. У межах цього блоку передбачено завантаження архіву з донавченою моделлю та керувальний елемент «Зареєструвати модель», який відповідає процедурі додавання моделі до локального реєстру для подальшого використання через інтерфейс без ручного введення шляхів. У тій же секції розміщено кнопку «Діагностика» та поле «Результат», які призначені для перевірки коректності завантаження моделі й ініціалізації лінгвістичного конвеєра синтаксичного аналізу. У нижній частині робочої області передбачено великий блок «Журнал системи (помилки, службові повідомлення)», який акумулює діагностичні дані й забезпечує прозорість при налагодженні та виявленні причин збоїв, пов’язаних із некоректним форматом моделі, відсутніми компонентами середовища або ресурсними обмеженнями. У цілому екран «Налаштування» забезпечує керованість і відтворюваність роботи системи, а також формує технічні передумови для коректного впровадження в різних обчислювальних середовищах і для підтримки декількох версій моделей у межах одного програмного інструменту.

На рисунку 3.9 подано прототип екрана «Про систему», який виконує довідково пояснювальну функцію та фіксує ключові відомості щодо архітектури, вхідних і вихідних даних, а також правил коректного трактування результатів аналізу.

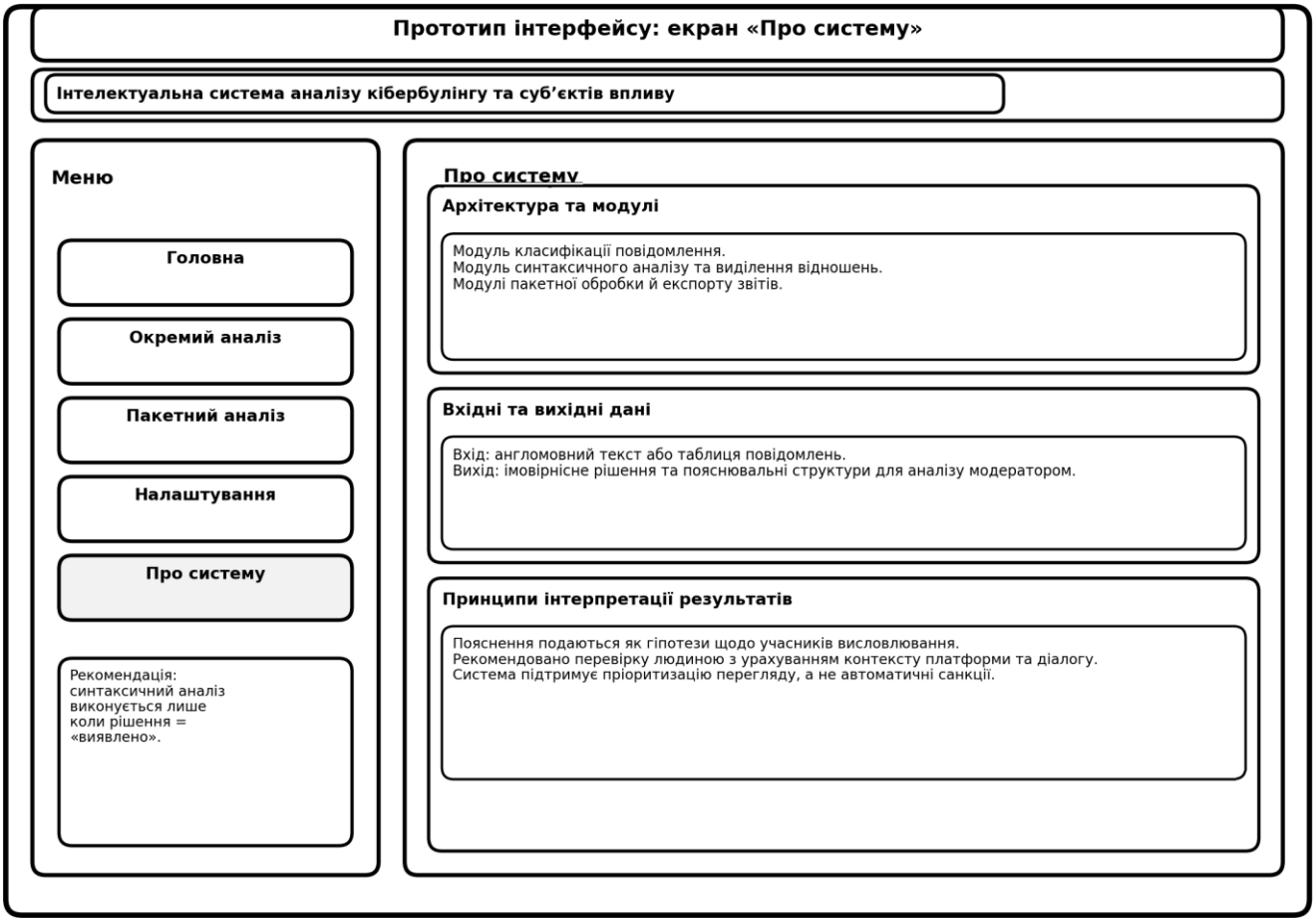


Рисунок 3.9 – Прототип інтерфейсу «Про систему»

Компоновка екрана є уніфікованою з іншими розділами інтерфейсу: у верхній частині розміщено службовий заголовок прототипу та шапку з назвою інтелектуальної системи, зліва наведено навігаційну панель «Меню» з переходами між основними екранами, а нижче розташовано короткий рекомендаційний блок, що відсилає до принципу умовного виконання синтаксичного аналізу. Таке рішення забезпечує сталість структури й дозволяє користувачу швидко орієнтуватися в системі незалежно від обраного режиму.

Права частина екрана містить основну інформаційну область «Про систему», яка структурована на три логічні секції. Перша секція «Архітектура та модулі» стисло описує внутрішню організацію програмного засобу через виокремлення функціональних компонентів: модуля класифікації повідомлень, модуля синтаксичного аналізу та виділення відношень між учасниками висловлювання, а також модуля пакетної обробки й експорту звітів. Така подача дозволяє користувачу пов’язати елементи інтерфейсу з відповідними етапами обчислювального конвеєра й розуміти, які саме підсистеми залучаються під час виконання конкретного сценарію.

Друга секція «Вхідні та вихідні дані» уточнює інформаційні межі системи: вхідними даними виступають англомовні тексти або таблиці повідомлень, тоді як вихід формують імовірнісне рішення моделі та пояснювальні структури, придатні для аналізу модератором. Фіксація формату входу й складу виходу є важливою для коректного використання системи у практичних умовах, оскільки дозволяє узгодити очікування користувача щодо результату та спрощує інтеграцію з процесами збирання даних або внутрішніми журналами модерації.

Третя секція «Принципи інтерпретації результатів» формалізує правила відповідального використання автоматизованих висновків. У цьому блоці підкреслюється, що пояснювальні структури не є юридичним чи остаточним доказом, а мають допоміжний характер і призначені для підтримки експертного перегляду з урахуванням контексту платформи та діалогу. Також акцентується, що система орієнтована на пріоритизацію перевірки й аналітичну підтримку, а не на автоматичне застосування санкцій. Таким чином, екран «Про систему» закріплює рамку коректної експлуатації, підвищує прозорість роботи програмного засобу та зменшує ризики некритичного трактування результатів у сценаріях модерації й моніторингу.

Виконано проєктування інтерфейсу користувача інтелектуальної системи аналізу кібербулінгу та суб’єктів впливу і сформовано прототипи основних екранів, що відображають ключові сценарії роботи: ознайомлення з призначенням і конвеєром обробки, інтерактивний аналіз одного повідомлення з пояснювальними результатами, пакетну обробку таблиць із експортом, а також конфігурування

моделі, порога прийняття рішення і процедур діагностики. Запропонована компоновка забезпечує послідовну навігацію, логічну послідовність, керованість параметрів і відтворюваність аналізу, а також підтримує практичні вимоги модерації та аналітики за рахунок умовної активації синтаксичного модуля і наявності журналювання для контролю помилок та експлуатаційних збоїв.

3.4 Вимоги до системи та сценарії використання у задачах модерації й аналітики

Проєктована інформаційна система призначена для автоматизованого аналізу англомовних текстових повідомлень з метою виявлення ознак кібербулінгу та формування інтерпретованого опису комунікативної ситуації, зокрема через встановлення ключових учасників взаємодії та характеру впливу. Вимоги до системи визначаються специфікою практичних процесів модерації й аналітики, де важливими є оперативність обробки, відтворюваність результатів, контроль налаштувань моделі, а також можливість пояснення отриманого висновку. Система має підтримувати два режими використання: індивідуальний аналіз повідомлення з формуванням деталізованого результату для конкретного кейсу та пакетний аналіз колекції повідомлень для цілей моніторингу, оцінювання ризиків і підготовки підсумкових звітів.

Функціональні вимоги формуються навколо повного конвеєра обробки даних. На вході система повинна приймати текстове повідомлення або набір повідомлень у форматах, придатних для операційної роботи, зокрема під час експорту з платформ соціальних мереж чи внутрішніх журналів. Далі забезпечується визначення класу повідомлення нейромережевою моделлю та розрахунок імовірнісної оцінки, що дозволяє керувати чутливістю залежно від контексту застосування. Для модераційних сценаріїв критичною є наявність керованого порогового механізму, оскільки допустимий рівень ризику хибнопозитивних та хибнонегативних рішень залежить від політик платформи й типу контенту. Водночас результат класифікації має супроводжуватися інтерпретаційним

компонентом, який формує структуроване подання виявлених відношень між учасниками висловлювання. Такий підхід забезпечує перехід від констатації факту токсичності до аналітично значущого опису того, хто є ініціатором впливу, на кого спрямовано повідомлення і якою лексико граматичною конструкцією цей вплив реалізовано. У системі також має бути реалізовано керування моделями, що включає вибір стандартної попередньо навченої моделі або завантаження донавчених варіантів, а також ведення реєстру моделей із фіксацією їхніх параметрів, версій та джерела походження. Це забезпечує відтворюваність експериментів і прозорість порівняння результатів у часі.

Нефункціональні вимоги визначають придатність системи до експлуатації в умовах реальних навантажень і обмежень ресурсів. Система має забезпечувати детермінованість ключових кроків виконання з фіксацією конфігурації запуску, що дозволяє повторно отримати співставні результати за однакових умов. Особливо важливою є вимога до пояснюваності, оскільки в модерації результати автоматизованого аналізу часто виступають підставою для подальших дій людини, включно з пріоритизацією перегляду або ініціюванням розслідування інциденту. У цьому контексті інтерпретаційні дані мають бути представлені у формі, що зручна для фахівця і допускає перевірку, наприклад через відображення структурованих відношень і агрегованих сутностей. Важливими є вимоги до надійності та керованої деградації: у випадку проблем із доступом до ресурсів, тимчасової недоступності моделей або помилок синтаксичного аналізу система повинна зберігати працездатність на базовому рівні, формуючи повідомлення про помилку в інтерфейсі та не блокуючи інші сценарії роботи. Окремою вимогою є захист даних, оскільки повідомлення можуть містити персональну або чутливу інформацію. Це зумовлює потребу в мінімізації зберігання сирого тексту, контролі доступу, а також у формуванні звітних артефактів таким чином, щоб вони були придатними для аналізу без надмірного розкриття приватних даних.

Сценарії використання системи у модерації передбачають підтримку швидкого аналізу окремого повідомлення та допоміжну інтерпретацію для ухвалення рішення людиною. Типовим є сценарій, коли модератор отримує

повідомлення або фрагмент діалогу й потребує оперативної оцінки ймовірності кібербулінгу. У такому режимі система повинна надавати короткий підсумок із числовою оцінкою, параметрами прийняття рішення та відображенням структури взаємодії між учасниками висловлювання. У випадках, коли система використовується як інструмент пріоритизації, важливою є можливість налаштування чутливості та використання інтерпретаційного аналізу лише для повідомлень, які мають підвищений ризик, що зменшує обчислювальні витрати і скорочує час обробки. За такого підходу забезпечується баланс між точністю та продуктивністю, а також знижується ризик перенавантаження модератора надлишковими деталями у випадках нейтральних повідомлень.

Сценарії використання в аналітиці орієнтовані на обробку масивів текстів з метою виявлення тенденцій, підготовки статистичних оглядів і формування доказової бази для управлінських рішень. У цьому режимі система повинна підтримувати пакетну обробку даних із можливістю обмеження обсягу, вибору полів ідентифікації та формування підсумкових таблиць результатів. Аналітична цінність підвищується за рахунок агрегування інформації щодо виявлених учасників і типових конструкцій впливу, що дозволяє ідентифікувати повторювані шаблони поведінки або класифікувати характер взаємодій у певній спільноті. Водночас у цьому сценарії важливо забезпечити зіставність результатів між різними періодами, що вимагає фіксації параметрів моделі, порогів прийняття рішення та версій компонентів синтаксичного аналізу. Таким чином, система має підтримувати не лише отримання результатів, а й повний контекст їх формування, що забезпечує наукову коректність інтерпретації та придатність результатів для подальшого використання в дослідженнях або звітності.

Загалом сукупність вимог і сценаріїв використання визначає систему як інструмент прикладної підтримки модераційних і аналітичних процесів, де ключовими характеристиками є керованість прийняття рішення, інтерпретованість результатів, відтворюваність запусків, надійність у умовах змінного навантаження та відповідальне поводження з користувацькими даними. Це створює підґрунтя для

подальшого проєктування архітектури реалізації, вибору механізмів розгортання та формалізації процедур тестування в умовах наближених до експлуатаційних.

3.5 Розгортання та експлуатація у хмарному середовищі, обмеження ресурсів і масштабованість

Розгортання системи аналізу кібербулінгу у хмарному середовищі визначається поєднанням двох обчислювально різних підзадач, а саме нейромережевої інференції та синтаксичного аналізу природної мови. Перша підзадача є ресурсомісткою з погляду обчислень на тензорних операціях і може вигравати від використання графічних прискорювачів, тоді як друга переважно навантажує центральний процесор і оперативну пам’ять. Тому у хмарній експлуатації доцільно розглядати систему як конвеєр із керованою конфігурацією ресурсів, де оптимізація досягається не лише вибором апаратної платформи, а й організацією виконання компонентів, кешуванням ініціалізованих моделей і контрольованим застосуванням інтерпретаційних процедур.

Типове розгортання в хмарі передбачає ізоляцію середовища виконання у вигляді контейнеризованого сервісу, де фіксуються версії бібліотек, моделі та конфігураційні параметри. Такий підхід зменшує ризики несумісності залежностей та забезпечує відтворюваність результатів між інстансами. Особливістю систем, що використовують трансформерні моделі, є значні витрати часу на первинне завантаження ваг, а також чутливість до параметрів пам’яті. У хмарній інфраструктурі це вимагає відокремлення етапу ініціалізації від етапу обробки запитів, використання прогріву сервісу та підтримки довгоживучого процесу, у якому моделі й мовні конвеєри зберігаються у пам’яті між запитами. Для середовищ із обмеженим часом життя сесії доцільним є зберігання артефактів у спільному сховищі та організація швидкого відновлення стану за допомогою локальних кешів або керованих образів середовища.

Обмеження ресурсів проявляються насамперед у трьох аспектах: оперативна пам’ять, пропускна здатність при пакетній обробці та час відповіді в інтерактивному

режимі. Для індивідуального аналізу ключовим показником є латентність, що залежить від розміру моделі, довжини тексту та режиму виконання, а також від того, чи активується синтаксичний модуль. Для пакетного аналізу визначальним стає throughput, тобто кількість повідомлень, оброблених за одиницю часу, а також стабільність роботи без деградації через накопичення стану або неконтрольоване зростання використання пам’яті. З огляду на це доцільним є застосування керованого обмеження обсягу вхідних даних на один запуск, контроль довжини текстів і реалізація політик тайм аутів для довгих або некоректно сформованих повідомлень. Окремо слід враховувати, що синтаксичний аналіз неформального тексту може мати змінну складність залежно від токенізації та структури речень, тому запуск цього етапу лише для повідомлень, які мають підвищену імовірність токсичності, є практично значущим механізмом зниження навантаження та стабілізації часу обробки.

Масштабованість системи у хмарі досягається переважно горизонтальним масштабуванням, коли однотипні екземпляри сервісу обробляють запити паралельно за балансувальником навантаження. Однак для моделей машинного навчання проста реплікація може бути економічно неефективною, оскільки кожен екземпляр потребує власного завантаження ваг моделі в пам’ять, що підвищує загальні витрати. Тому при плануванні масштабування доцільно відокремлювати сервіс інференції від сервісу інтерфейсу, а також розглядати асинхронну обробку пакетних задач. У такій організації інтерактивні запити користувача обслуговуються окремим контуром із пріоритетом низької латентності, тоді як пакетні запуски можуть виконуватися у фоновому контурі з чергою задач і контрольованим використанням ресурсів. Це дозволяє уникати конкуренції за ресурси між різними режимами роботи та забезпечувати прогнозовану якість сервісу під час пікових навантажень.

Окремого розгляду потребує вибір між центральним процесором і графічним прискорювачем. Для малих обсягів даних або моделей компактного розміру виконання на центральному процесорі може бути достатнім і економічно виправданим, особливо якщо критичним є мінімізація вартості експлуатації.

Використання графічного прискорювача доцільне у випадках інтенсивної інференції, довших текстів або одночасного обслуговування багатьох користувачів, коли виграш у продуктивності компенсує вартість ресурсу. Водночас навіть за наявності графічного прискорювача синтаксичний компонент часто залишається процесорно залежним, що зумовлює потребу у збалансованому виділенні обчислювальних ресурсів на рівні вузла та у виборі конфігурацій з достатнім обсягом оперативної пам’яті. З технічної точки зору суттєвий ефект дає оптимізація організації обробки, зокрема повторне використання ініціалізованих об’єктів моделі, мінімізація повторних завантажень та уникнення дублювання конвеєра між запитами.

Експлуатаційна придатність системи у хмарі пов’язана також із моніторингом і керуванням якістю. Стабільна робота потребує збору метрик латентності, частоти помилок, використання пам’яті та процесорного часу, а також журналювання контексту запуску, включно з версією моделі та конфігурацією порогового рішення. Такий підхід забезпечує можливість діагностики деградації якості, наприклад при зміні домену вхідних даних або при оновленні залежностей. З позицій безпеки та приватності в хмарному розгортанні має бути передбачено контроль доступу до інтерфейсу, розмежування прав на завантаження даних і моделей, а також політики обмеженого зберігання й очищення тимчасових файлів. Це зменшує ризики витоку даних і відповідає вимогам відповідального поводження з користувацьким контентом.

Таким чином, розгортання системи у хмарному середовищі повинно враховувати гетерогенність обчислювальних потреб компонентів, обмеження пам’яті й пропускної здатності, а також необхідність масштабування під різні режими експлуатації. Поєднання контейнеризованого відтворюваного середовища, керованого виконання інтерпретаційних етапів, розмежування контурів інтерактивної та пакетної обробки, а також системного моніторингу формує технологічні передумови для стабільної роботи, прогнозованої продуктивності й економічно обґрунтованого масштабування в умовах змінного навантаження.

Висновки до розділу 3

У розділі виконано проєктування інтелектуальної системи багаторівневого виявлення суб’єктів впливу кібербулінгу, у межах якого обґрунтовано інструментальну основу реалізації, визначено склад і взаємодію програмних компонентів, а також формалізовано вимоги до інтерфейсу, сценаріїв застосування та умов експлуатації в хмарному середовищі. Запропоноване проєктне рішення підтримує як інтерактивний режим роботи модератора з одиничними повідомленнями, так і аналітичний режим обробки масивів даних із формуванням відтворюваних артефактів, придатних для подальшої інтерпретації та звітності.

Обґрунтовано вибір засобів розробки, які забезпечують сумісність із трансформерними моделями, синтаксичним аналізом та інтерфейсною частиною системи. Визначено, що використання Python як базової мови реалізації є методично доцільним завдяки наявності стабільних бібліотек для NLP і глибинного навчання, а хмарне середовище Google Colab дозволяє відтворювано виконувати експерименти з мінімальними витратами на налаштування інфраструктури. Обраний набір інструментів формує практично придатну основу для прототипування, тестування й подальшого масштабування рішення.

Спроєктовано складові системи та описано їх взаємодію як послідовного конвеєра з керуванням через центральний контролер. Встановлено, що архітектура реалізує два рівні обробки: нейромережеву класифікацію як базовий рівень і синтаксико-семантичне структурування як пояснювальний рівень, який активується умовно. Формалізовано розгалуження виконання за пороговим правилом, що забезпечує кероване залучення ресурсомістких процедур і зменшує ризик застосування інтерпретацій до нейтральних повідомлень. Показано, що результат системи є комбінованим: поєднує імовірнісне рішення моделі з інтерпретованими структурами взаємодії, що дозволяє переходити від констатації токсичності до пояснення ролей учасників висловлювання. Розроблені діаграми варіантів використання, активності та послідовностей уточнюють як поведінку системи в

типових сценаріях, так і динаміку обміну даними між компонентами, що підвищує проєктну визначеність та однозначність реалізації.

Виконано проєктування інтерфейсу користувача та сформовано прототипи основних екранів, орієнтовані на прикладні потреби модерації й аналітики. Узгоджено уніфіковану структуру навігації, у межах якої користувач має сталу ліву панель меню та логічно згруповані інформаційні блоки у робочій області. Прототипи екранів «Головна», «Окремий аналіз», «Пакетний аналіз», «Налаштування» та «Про систему» відображають повний цикл взаємодії: від ознайомлення з призначенням і конвеєром обробки до отримання результатів, експорту артефактів і керування конфігурацією моделі. Підкреслено, що мінімалістична візуальна організація та послідовність елементів керування зменшують когнітивне навантаження оператора, забезпечують швидке виконання типових дій і підтримують контрольовану інтерпретацію результатів через структуроване представлення відношень і журналювання.

Сформовано вимоги до системи та описано сценарії використання в задачах модерації й аналітики, що дозволяє оцінювати проєкт не лише за функціональною повнотою, а й за експлуатаційною придатністю. Визначено, що система має забезпечувати прийом текстів і табличних наборів повідомлень, обчислення імовірнісної оцінки та прийняття рішення з керованим порогом, а також формування пояснювального шару у вигляді структурованих відношень між учасниками висловлювання. Обґрунтовано необхідність керування моделями та ведення реєстру конфігурацій для відтворюваності й порівнянності результатів у часі. Окремо уточнено нефункціональні вимоги, критичні для реальної експлуатації: детермінованість ключових кроків, пояснюваність для підтримки людини в контурі ухвалення рішення, надійність із керованою деградацією при помилках компонентів, а також вимоги приватності та мінімізації роботи з сирими даними. Сформульовані сценарії показують, що система придатна як для швидкого аналізу окремих інцидентів модератором, так і для пакетної аналітики з отриманням зведених таблиць і контексту їх формування.

Розглянуто принципи розгортання та експлуатації системи у хмарному середовищі з урахуванням ресурсних обмежень і вимог масштабованості. Показано, що обчислювальне навантаження є гетерогенним: трансформерна інференція є переважно тензорною та виграє від графічного прискорення, тоді як синтаксичний аналіз є більш процесорно і пам’яттєзалежним, що потребує збалансованого виділення ресурсів. Обґрунтовано доцільність контейнеризованого підходу для фіксації залежностей і забезпечення відтворюваності, а також необхідність збереження ініціалізованих моделей у пам’яті між запитами для зменшення латентності та уникнення повторних завантажень. Визначено ключові обмеження експлуатації, пов’язані з оперативною пам’яттю, пропускною здатністю при пакетній обробці та часом відповіді в інтерактивному режимі, і показано роль умовної активації синтаксичного модуля як механізму стабілізації продуктивності. Описано принципи масштабування, зокрема горизонтальне розширення з урахуванням вартості реплікації моделей, доцільність розділення контурів інтерактивної та пакетної обробки, а також вимоги до моніторингу, журналювання й контролю доступу, що забезпечує керованість сервісу та відповідальне поводження з даними.

Таким чином, розділ 3 забезпечив завершене проєктне обґрунтування інтелектуальної системи на рівні інструментів реалізації, архітектурної логіки, інтерфейсного представлення, експлуатаційних вимог і хмарної придатності. Сукупність отриманих результатів створює методичну та інженерну основу для переходу від прототипу до реалізації, а також для подальшого тестування якості, валідації переносимості на різних доменах і інтеграції у процеси модерації та аналітичного моніторингу.

РОЗДІЛ 4 Експериментальне дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей

4.1 Програмна структура компонентів інтелектуальної системи

Програмна структура інтелектуальної системи у вигляді діаграми класів наведена на рисунку 4.1.

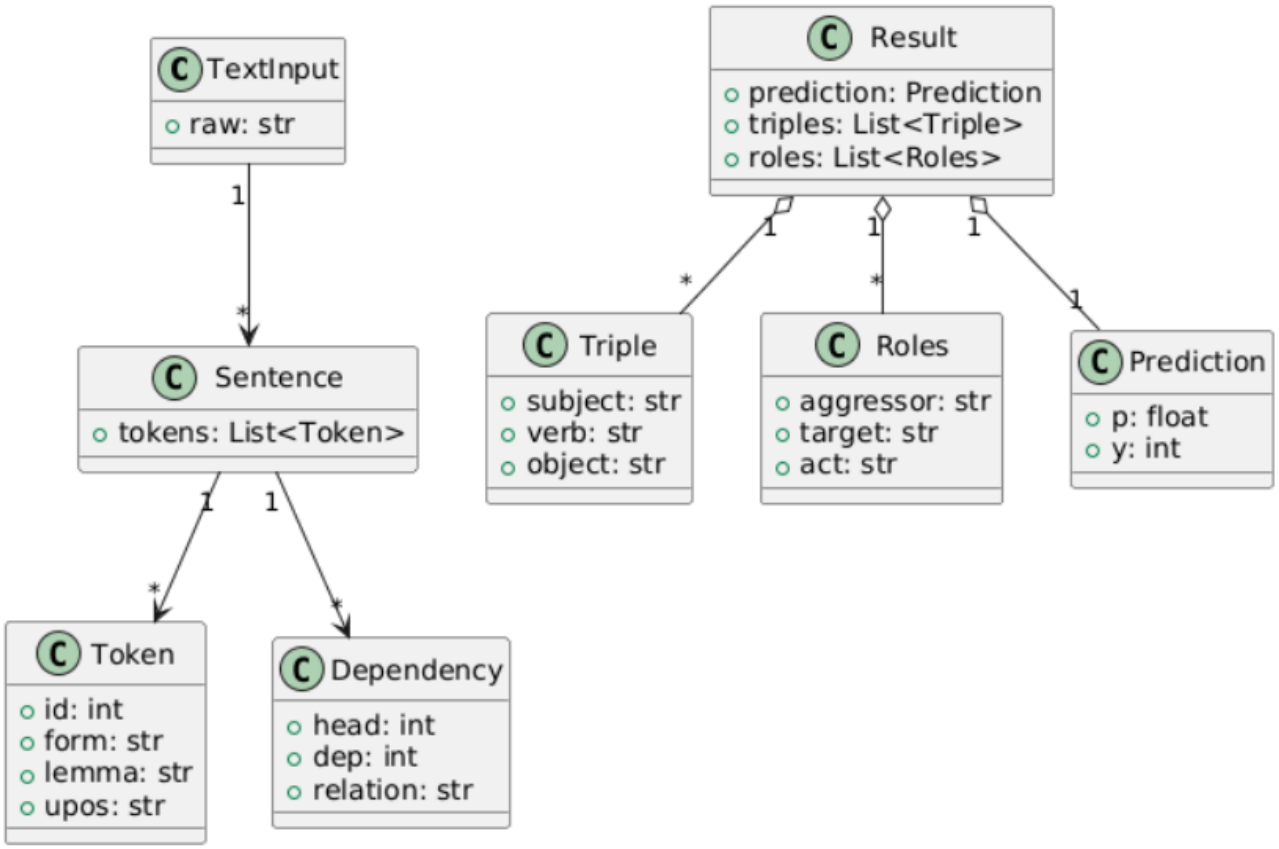


Рисунок 4.1 – Діаграма класів

Подана діаграма класів задає формальну модель даних системи й фіксує, які сутності зберігаються під час обробки та як вони пов’язані між собою. Вхідний текст інкапсульовано в об’єкті TextInput, що містить сирцеве повідомлення та є єдиним джерелом для подальших структур. Із нього породжується множина Sentence; кардинальність «один до багатьох» відображає сегментацію повідомлення на речення. Кожне речення асоціює в собі послідовність токенів і набір синтаксичних дуг. Клас Token зберігає позиційний ідентифікатор, поверхневу

форму, лему та частиномовний тег, а Dependency фіксує орієнтований зв’язок у термінах голови, залежного й типу відношення; зв’язки один-до-багатьох між Sentence і цими класами відображають дерево залежностей у межах речення.

Семантичний рівень представлено класами Triple та Roles. Перша структура моделює елементарну дію у форматі «підмет - предикат - об’єкт» і є безпосереднім результатом трансформації синтаксичних залежностей у рольові зв’язки. Друга структура відображає інтерпретацію впливу в домені кібербулінгу, де з трійки виділяються агресор, адресат і акт впливу. Така декомпозиція дозволяє розмежувати лінгвістичну реконструкцію від предметно-орієнтованої інтерпретації та забезпечує прозорість переходу від граматичних відношень до комунікативних ролей.

Показник класифікації зберігається в класі Prediction, який містить імовірність наявності кібербулінгу та бінарне рішення після порогування. Клас Result агрегує всі вихідні артефакти обробки: посилання на Prediction, колекцію трійок і колекцію інтерпретованих ролей. Позначення заповненої ромбової асоціації між Result і підлеглими об’єктами означає композицію: життєвий цикл трійок, ролей і прогнозу підпорядкований життєвому циклу єдиного результату запиту. Така організація даних узгоджується з конвеєром системи: від TextInput через синтаксичні структури до Triple і Roles, після чого все зводиться в Result, який і повертається інтерфейсу користувача.

Зображена схема (рисунок 4.2) передає архітектурну організацію середовища, у якому виконується система. У центрі перебуває платформа Google Colab, що виступає контейнером для всіх компонентів розробки й запуску. Усередині цього середовища розташовано Python Runtime – фактичне оточення, де виконуються модулі системи. До нього входять окремі блоки, кожен з яких відповідає за певну функціональність: Gradio App реалізує інтерфейс взаємодії, Controller координує етапи обробки, Transformers забезпечує класифікацію на основі попередньо навчених моделей, Stanza відповідає за синтаксико-семантичний аналіз, а pandas використовується для структурування та проміжного опрацювання даних.

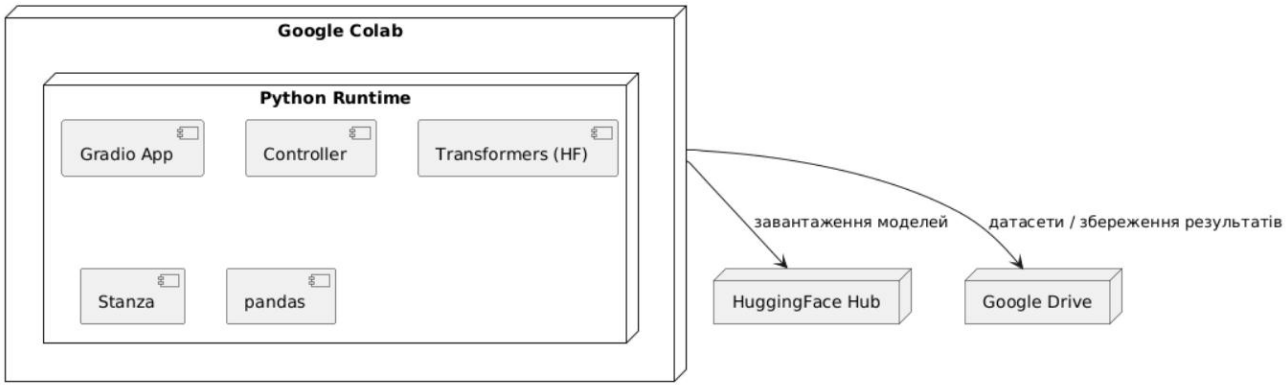


Рисунок 4.2 – Архітектурна організація середовища

Середовище виконання не ізольоване: воно взаємодіє із зовнішніми ресурсами. HuggingFace Hub використовується як джерело для завантаження моделей, що підключаються безпосередньо до Python Runtime. Google Drive виконує функції сховища для корпусів, допоміжних даних і результатів, які можуть зберігатися або переноситися між сесіями. Таким чином, схема фіксує не лише програмні компоненти, а й залежності від зовнішніх сервісів, що забезпечують відтворюваність, доступ до моделей і підтримку даних.

4.2 Особливості розробки прикладних компонентів інтелектуальної системи

Нижче наведено набір базових алгоритмів, які відображають ключові прикладні компоненти системи, а також логіку їх взаємодії під час обробки повідомлень у режимі окремого та пакетного аналізу. Псевдокоди подано у стислому вигляді, і вони орієнтовано на відтворення проєктної логіки, а не конкретної реалізації певною бібліотекою.

Алгоритм 4.1. Ініціалізація середовища та підготовка каталогів:

```
процедура ІНІЦІАЛІЗУВАТИ_СЕРЕДОВИЩЕ()
    якщо НЕ існує(КАТАЛОГ_STANZA) тоді
        створити_каталог(КАТАЛОГ_STANZA)
    кінець якщо

    якщо НЕ існує(КАТАЛОГ_РЕЄСТРУ_МОДЕЛЕЙ) тоді
        створити_каталог(КАТАЛОГ_РЕЄСТРУ_МОДЕЛЕЙ)
```

```
    кінець якщо

    налаштувати_кодування("utf-8")
    ініціалізувати_журналювання()

    спробувати
        завантажити_мовні_ресурси_STANZA(мова="en", каталог=КАТАЛОГ_STANZA)
    зловити помилка
        записати_в_журнал("Не вдалося завантажити ресурси Stanza", помилка)
        підняти помилка
    кінець спроби
кінець процедури
```

У цьому алгоритмі формалізовано стартові дії, потрібні для відтворюваного запуску. Практично важливо одразу створювати каталоги для мовних ресурсів і реєстру моделей, оскільки ці артефакти повторно використовуються між сесіями. Додатково ініціалізується журналювання для фіксації помилок середовища, що спрощує експлуатаційну діагностику.

Алгоритм 4.2. Обирання джерела моделі та валідація параметрів:

```
функція ВИЗНАЧИТИ_МОДЕЛЬ(режим_моделі, hf_ід, локальний_шлях) повертає рядок
    якщо режим_моделі = "Стандартна" тоді
        якщо порожній(hf_ід) тоді
            підняти ПОМИЛКА("Не задано ідентифікатор моделі HuggingFace")
        кінець якщо
        повернути обрізати(hf_ід)
    інакше
        якщо порожній(локальний_шлях) тоді
            підняти ПОМИЛКА("Не задано шлях до локальної моделі")
        кінець якщо
        якщо НЕ є_каталогом(локальний_шлях) тоді
            підняти ПОМИЛКА("Локальна модель не знайдена або шлях некоректний")
        кінець якщо
        повернути обрізати(локальний_шлях)
    кінець якщо
кінець функції
```

Алгоритм задає однозначне правило вибору нейромережі: або стандартна модель за ідентифікатором, або локальна модель за шляхом. Валідація параметрів виконується до запуску інференції, щоб уникати часткових виконань і неузгоджених станів, особливо в інтерактивному інтерфейсі.

Алгоритм 4.3. Реєстрація локальної моделі з архіву та нормалізація структури:

```
процедура ЗАРЕЄСТРУВАТИ_МОДЕЛЬ_ZIP(архів_zip, назва)
```

```
    якщо архів_zip = НІ тоді
        повернути "Архів не надано"
    кінець якщо

    якщо порожній(назва) тоді
        назва = "model_" + поточний_час()
    кінець якщо

    ціль = КАТАЛОГ_РЕЄСТРУ_МОДЕЛЕЙ + "/" + назва
    якщо існує(ціль) тоді
        видалити_каталог(ціль)
    кінець якщо
    створити_каталог(ціль)

    розпакувати(архів_zip, у_каталог=ціль)

    якщо кількість_елементів(ціль) = 1 і елемент_є_каталогом тоді
        внутрішній = перший_елемент(ціль)
        перемістити_вміст(внутрішній, у_ціль=ціль)
        видалити_каталог(внутрішній)
    кінець якщо

    повернути "Модель зареєстровано: " + ціль
кінець процедура
```

Алгоритм відображає прикладний механізм ведення реєстру моделей. Ключовим є крок нормалізації структури, коли архів містить один кореневий каталог: вміст переноситься на верхній рівень, щоб уніфікувати подальше завантаження моделі незалежно від способу пакування.

Алгоритм 4.4. Нормалізація виходу класифікатора та обчислення оцінки класу:

```
функція НОРМАЛІЗУВАТИ_КЛАС(мітка, оцінка) повертає (p_кібербулінг, мітка_сир)
    мітка_сир = верхній_регістр(мітка)
    якщо мітка_сир у {"OFF", "OFFENSIVE", "LABEL_1"} тоді
        p_кібербулінг = оцінка
    інакше якщо мітка_сир у {"NOT", "LABEL_0"} тоді
        p_кібербулінг = 1.0 - оцінка
    інакше
        p_кібербулінг = оцінка
    кінець якщо
    повернути (p_кібербулінг, мітка_сир)
кінець функції
```

Цей алгоритм усуває неоднозначність інтерпретації виходу різних моделей, де позитивний клас може бути закодований різними мітками. Окремо враховано випадок, коли модель повертає оцінку для негативного класу: тоді ймовірність кібербулінгу обчислюється як доповнення до одиниці. Саме цей крок часто є джерелом некоректних висновків у прикладних прототипах.

Алгоритм 4.5. Прийняття рішення за міткою або за порогом, з керованим правилом:

```
функція ПРИЙНЯТИ_РІШЕННЯ(p_кібербулінг, мітка_сир, τ, режим_правила) повертає логічне
    якщо режим_правила = "за_міткою" тоді
        якщо мітка_сир у {"OFF", "OFFENSIVE", "LABEL_1"} тоді
            повернути ІСТИНА
        інакше
            повернути ХИБНІСТЬ
        кінець якщо
    інакше
        якщо p_кібербулінг >= τ тоді
            повернути ІСТИНА
        інакше
            повернути ХИБНІСТЬ
        кінець якщо
    кінець якщо
кінець функції
```

Алгоритм формалізує два практичні підходи до ухвалення бінарного висновку. Правило за міткою корисне, коли вихідна модель вже оптимізована під фіксовану границю рішення. Правило за порогом забезпечує керованість чутливості, що критично для модерації з різними політиками ризику. Важливо, що обидва правила спираються на вже нормалізовану інтерпретацію класу.

Алгоритм 4.6. Керована активація синтаксичного аналізу та формування результату:

```
функція АНАЛІЗУВАТИ_ПОВІДОМЛЕННЯ(текст, модель, τ, режим_правила, gating) повертає РЕЗУЛЬТАТ
    якщо порожній(обрізати(текст)) тоді
        повернути РЕЗУЛЬТАТ(статус="порожній_ввід")
    кінець якщо

    (мітка, оцінка) = ЗАПУСТИТИ_КЛАСИФІКАТОР(модель, текст)
    (p, мітка_сир) = НОРМАЛІЗУВАТИ_КЛАС(мітка, оцінка)
    рішення = ПРИЙНЯТИ_РІШЕННЯ(p, мітка_сир, τ, режим_правила)

    якщо gating = ІСТИНА і рішення = ХИБНІСТЬ тоді
        повернути РЕЗУЛЬТАТ(рішення=ХИБНІСТЬ, p=p, мітка=мітка_сир, трійки=порожньо)
    кінець якщо

    залежності = СИНТАКСИЧНИЙ_РОЗБІР_STANZA(текст)
    трійки = ВИТЯГНУТИ_ВІДНОШЕННЯ(залежності)
    повернути РЕЗУЛЬТАТ(рішення=рішення, p=p, мітка=мітка_сир, трійки=трійки)
кінець функції
```

Алгоритм відображає центральну логіку багаторівневої обробки. Перший рівень дає ймовірнісний висновок, другий рівень формує пояснювальні структури лише тоді, коли це обґрунтовано рішенням або коли gating вимкнено. Такий підхід зменшує навантаження та підвищує якість інтерпретації, оскільки структурні ролі не обчислюються для повідомлень, де відсутні ознаки кібербулінгу.

Алгоритм 4.7. Витягування відношень підмет, дієслово, об’єкт або комплемент:

```
        функція ВИТЯГНУТИ_ВІДНОШЕННЯ(залежності) повертає список_трійок
            трійки = порожній_список()

            для кожного речення у залежності.речення зробити
                предикати = ЗНАЙТИ_ПРЕДИКАТИ(речення)           # дієслова та предикативні
конструкції
                адресати = ЗНАЙТИ_ЗВЕРТАННЯ(речення)         # вокативні форми, якщо є
                для кожного pred у предикати зробити
                    підмети = ЗІБРАТИ_ПІДМЕТИ(речення, pred)
                    якщо порожньо(підмети) і НЕ порожньо(адресати) тоді
                        підмети = адресати
                    кінець якщо

                    цілі = ЗІБРАТИ_ОБЄКТИ_АБО_КОМПЛЕМЕНТИ(речення, pred)
                    якщо pred.тип = "копула" тоді
                        дієслово = "be"
                        комплемент = РОЗШИРИТИ_ПРЕДИКАТИВ(речення, pred)
                        цілі = {комплемент}
                    інакше
                        дієслово = pred.лема
                    кінець якщо
                    якщо порожньо(підмети) тоді підмети = {"unknown"} кінець якщо
                    якщо порожньо(цілі) тоді цілі = {"unknown"} кінець якщо
                    додати_декартово(трійки, підмети, дієслово, цілі)
                кінець для
            кінець для
            повернути трійки
        кінець функції
```

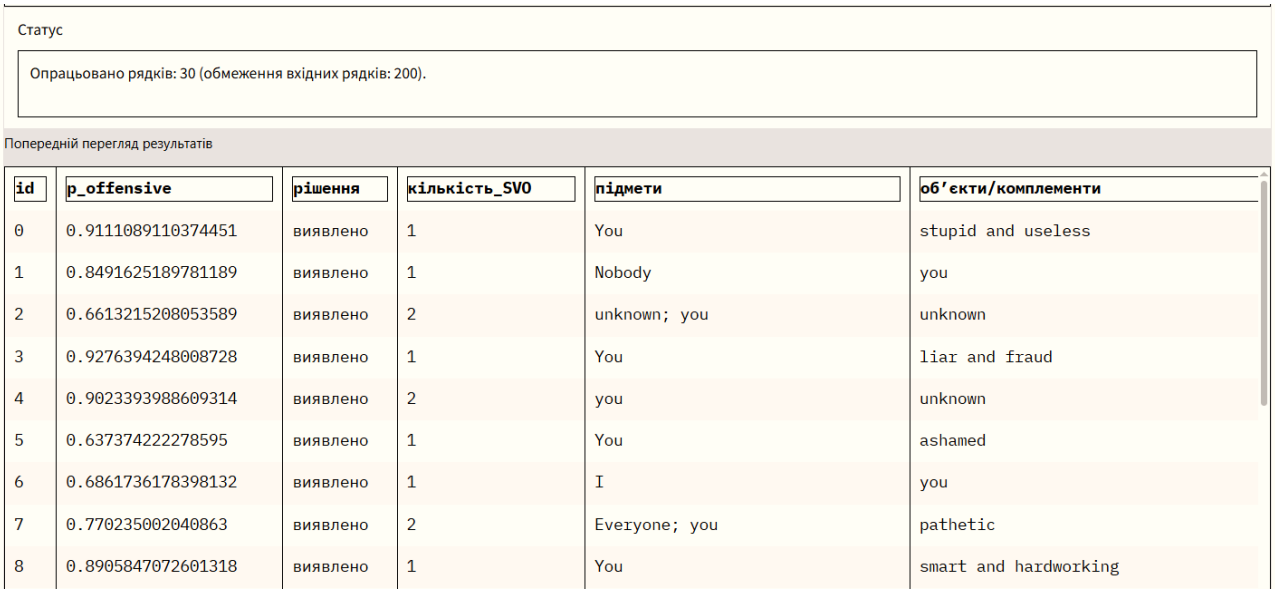
Алгоритм узагальнює процедуру переходу від залежнісного розбору до пояснювальних трійок, що використовуються для встановлення учасників впливу. Важливою є окрема обробка предикативних конструкцій з копулою, де формально предикат може бути прикметником або іменником, але логічно виконує роль присудка. Додатково враховано звертання як резервний механізм відновлення підмета у коротких репліках. Приклад виконання витягування відношень наведено на рисунку 4.3.

Рисунок 4.3 – Приклад виконання аналізу

Алгоритм 4.8. Пакетний аналіз таблиці та експорт результатів:

```
процедура ПАКЕТНИЙ_АНАЛІЗ(файл, колонка_тексту, колонка_id, τ, режим_правила, gating, максимум_рядків)
    таблиця = ЗЧИТАТИ_CSV_АБО_TSV(файл)
    якщо НЕ існує_колонка(таблиця, колонка_тексту) тоді
        підняти ПОМИЛКА("Колонка з текстом відсутня")
    кінець якщо

    результати = порожній_список()
    лічильник = 0

    для кожного рядок у перші_N(таблиця, максимум_рядків) зробити
        текст = обрізати(рядок[колонка_тексту])
        якщо порожній(текст) тоді продовжити кінець якщо

        res = АНАЛІЗУВАТИ_ПОВІДОМЛЕННЯ(текст, модель, τ, режим_правила, gating)

        id = якщо існує_колонка(таблиця, колонка_id) тоді рядок[колонка_id] інакше лічильник кінець якщо
        додати(результати, СФОРМУВАТИ_РЯДОК_ВИХОДУ(id, res))

        лічильник = лічильник + 1
    кінець для

    ЗБЕРЕГТИ_CSV(результати, "результати_пакетного_аналізу.csv")
    повернути "Опрацьовано рядків: " + лічильник
кінець процедура
```

Пакетний алгоритм (приклад виконання на рисунку 4.4) орієнтований на відтворювану обробку вибірок, що надходять у табличному форматі, та на формування вихідного файлу для подальшої аналітики. Валідація назви колонки з текстом виконується до запуску циклу, що запобігає частковим результатам. Додатково підтримується ідентифікація рядка або за окремою колонкою, або за індексом, що спрощує інтеграцію з різними структурами датасетів.


Статус

Опрацьовано рядків: 30 (обмеження вхідних рядків: 200).

Попередній перегляд результатів

id	p_offensive	рішення	кількість_SVO	підмети	об'єкти/комплементи
0	0.9111089110374451	виявлено	1	You	stupid and useless
1	0.8491625189781189	виявлено	1	Nobody	you
2	0.6613215208053589	виявлено	2	unknown; you	unknown
3	0.9276394248008728	виявлено	1	You	liar and fraud
4	0.9023393988609314	виявлено	2	you	unknown
5	0.637374222278595	виявлено	1	You	ashamed
6	0.6861736178398132	виявлено	1	I	you
7	0.770235002040863	виявлено	2	Everyone; you	pathetic
8	0.8905847072601318	виявлено	1	You	smart and hardworking

Рисунок 4.4 – Приклад пакетного виконання аналізу

Отже, формалізовано ключові прикладні компоненти інтелектуальної системи у вигляді псевдокодів, що відображають повний цикл обробки повідомлень від ініціалізації середовища й вибору моделі до прийняття класифікаційного рішення, керованої активації синтаксичного аналізу, виділення структур “підмет, дієслово, об’єкт або комплемент” та пакетного опрацювання таблиць із експортом результатів. Запропонований набір алгоритмів забезпечує однозначність реалізації критичних процедур, зокрема нормалізації міток і оцінок різних моделей та вибору правила ухвалення рішення, що знижує ризик помилкових висновків у прототипі. Одночасно виділення окремих алгоритмів для реєстру моделей, журналювання й валідації вхідних параметрів підвищує відтворюваність експериментів і керованість експлуатації, а умовне застосування структурного аналізу дозволяє збалансувати пояснюваність результатів із обмеженнями обчислювальних ресурсів у практичних сценаріях модерації та аналітики.

4.3 Прикладне тестування реалізації інтелектуальної системи

Прикладне тестування реалізації інтелектуальної системи виконано у формі модульних перевірок, спрямованих на підтвердження коректності ключових елементів прикладної логіки: інтерпретації результатів нейромережевої

класифікації, вибору джерела моделі, а також роботи обгорток, що формують вихідні артефакти для інтерфейсу. Сукупність тестів орієнтована на відтворюваний запуск у стандартному середовищі Python та перевіряє типові і критичні для експлуатації сценарії, зокрема обробку коректних і помилкових конфігурацій, гейтінг синтаксичного аналізу, формування табличних результатів і експорт пакетних результатів.

У тесті test_offensive_probability_negative_label перевіряється коректність перетворення виходу класифікатора для негативної мітки. Якщо модель повертає мітку типу NOT з певним значенням score, система повинна узгоджено інтерпретувати результат як низьку ймовірність наявності кібербулінгу (через комплементарність 1 мінус score) та повернути коректну пару значень “ймовірність, мітка”, придатну для подальшого порогового рішення та відображення в інтерфейсі.

У тесті test_offensive_probability_positive_label перевіряється симетричний позитивний випадок. За умови, що класифікатор повертає мітку типу OFF та оцінку score, система повинна трактувати цю оцінку без інверсій як ймовірність кібербулінгу й повернути її разом з міткою, забезпечуючи узгодженість числового значення з семантикою класу.

У тесті test_resolve_model_id_local_bad_path перевіряється валідація некоректного шляху до локальної, донавченої моделі. Якщо заданий шлях відсутній або не відповідає очікуваній структурі (не є директорією моделі), функція вибору моделі повинна завершуватися контрольованою помилкою (винятком). Така поведінка є принциповою для надійності, оскільки запобігає запуску інференції з невалідною конфігурацією та зменшує ризик неочевидних збоїв на пізніших етапах обробки.

У тесті test_resolve_model_id_local_ok перевіряється коректний сценарій вибору локальної моделі. Для тимчасово створеної директорії, що імітує коректний шлях до локального артефакту моделі, функція повинна повернути саме цей шлях як ідентифікатор моделі. Тест підтверджує, що локальний режим підтримується без додаткових залежностей від зовнішніх репозиторіїв та що система здатна працювати з донавченими версіями.

У тесті test_resolve_model_id_standard_missing перевіряється обов’язковість ідентифікатора моделі для стандартного режиму використання HuggingFace. Якщо поле ідентифікатора порожнє або містить лише пробіли, функція повинна формувати помилку конфігурації через виняток. Це забезпечує явність вимог до налаштувань і запобігає спробам ініціалізації pipeline без конкретної моделі.

У тесті test_resolve_model_id_standard_ok перевіряється коректність опрацювання ідентифікатора стандартної моделі, включно з нормалізацією введення. Рядок ідентифікатора з зайвими пробілами має бути очищений та повернений як валідний ідентифікатор, що підтверджує стійкість системи до типових помилок введення параметрів у користувацькому інтерфейсі.

У тесті test_analyze_single_empty_text перевіряється поведінка обгортки аналізу одиничного повідомлення для порожнього або неінформативного вводу. Система повинна коректно обробляти цей випадок без аварійного завершення, повертати контрольований результат для інтерфейсу та формувати зрозуміле службове повідомлення або стан, який унеможливлює помилкову інтерпретацію “порожнього” аналізу як реального висновку моделі.

У тесті test_analyze_single_extracts_relations_when_positive перевіряється логіка повного конвеєра для одиничного повідомлення за позитивного рішення класифікатора. За умови, що класифікатор повертає позитивний результат, повинні активуватися етапи синтаксичного аналізу й витягування відношень, а вихід має містити сформовані табличні структури (відношення та агреговані ролі), придатні для подання в інтерфейсі та експорту.

У тесті test_analyze_single_gating_skips_syntax перевіряється гейтінг інтерпретаційного шару. За негативного рішення класифікатора система повинна пропускати синтаксичний аналіз і не формувати відношення, що зменшує обчислювальні витрати та запобігає породженню пояснювальних структур там, де відсутні підстави для інтерпретації як кібербулінгу.

У тесті test_batch_analysis_outputs_csv перевіряється працездатність пакетного режиму з формуванням вихідного файлу результатів. За умови імітації роботи класифікатора на множині рядків система повинна сформувати узгоджений

табличний вихід і забезпечити експорт результатів у форматі CSV, що підтверджує придатність реалізації до аналітичних сценаріїв обробки масивів повідомлень і подальшої інтеграції з зовнішніми процесами.

На рисунку 4.5 наведено звіт з проведеного юніт-тестування.

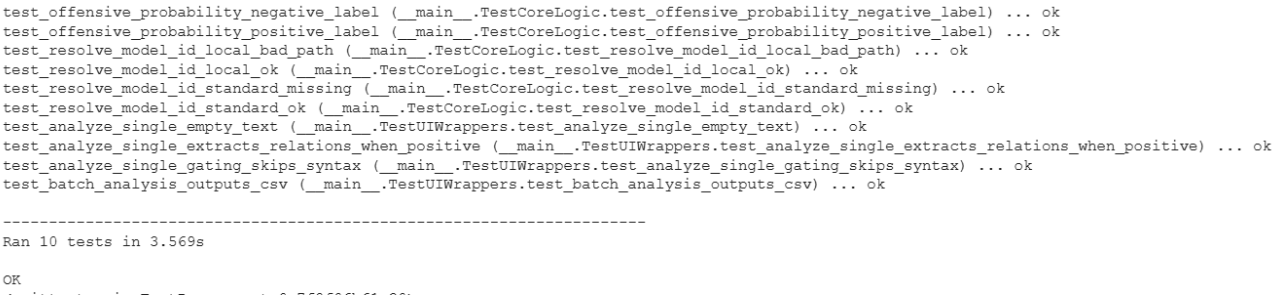


Рисунок 4.5 – Успішне виконання юніт-тестування

Отримані результати модульного тестування підтверджують коректність реалізації ключових прикладних компонентів інтелектуальної системи на рівні критичних для експлуатації сценаріїв. Успішне проходження тестів засвідчує, що система керовано обробляє помилкові налаштування, коректно реалізує механізм гейтінгу синтаксичного аналізу, формує структуровані табличні представлення та забезпечує експорт результатів, що в сукупності підвищує відтворюваність, надійність і практичну придатність реалізації для задач модерації й аналітики.

4.4 Особливості використання інтелектуальної системи

Використання інтелектуальної системи виявлення кібербулінгу та пояснення суб’єктів впливу з позиції користувача визначається тим, що система поєднує два різні за природою результати: класифікаційний висновок щодо наявності ознак кібербулінгу та інтерпретаційне подання взаємодії учасників повідомлення на основі синтаксичного аналізу. Для користувача це означає, що робота із системою не зводиться до отримання бінарної відповіді, а передбачає коректне читання показника впевненості моделі, розуміння умов, за яких формується пояснювальний шар, і співвіднесення цих даних із контекстом комунікації. Практична цінність

системи проявляється тоді, коли користувач використовує її як інструмент підтримки рішення: спочатку отримує швидку оцінку ризику, а потім, у разі потреби, переходить до структурованого пояснення, яке допомагає з’ясувати, хто є ініціатором висловлювання, на кого спрямовано вплив і якою мовною конструкцією він реалізований. Приклад визначення кібербулінгу наведено на рисунку 4.6.

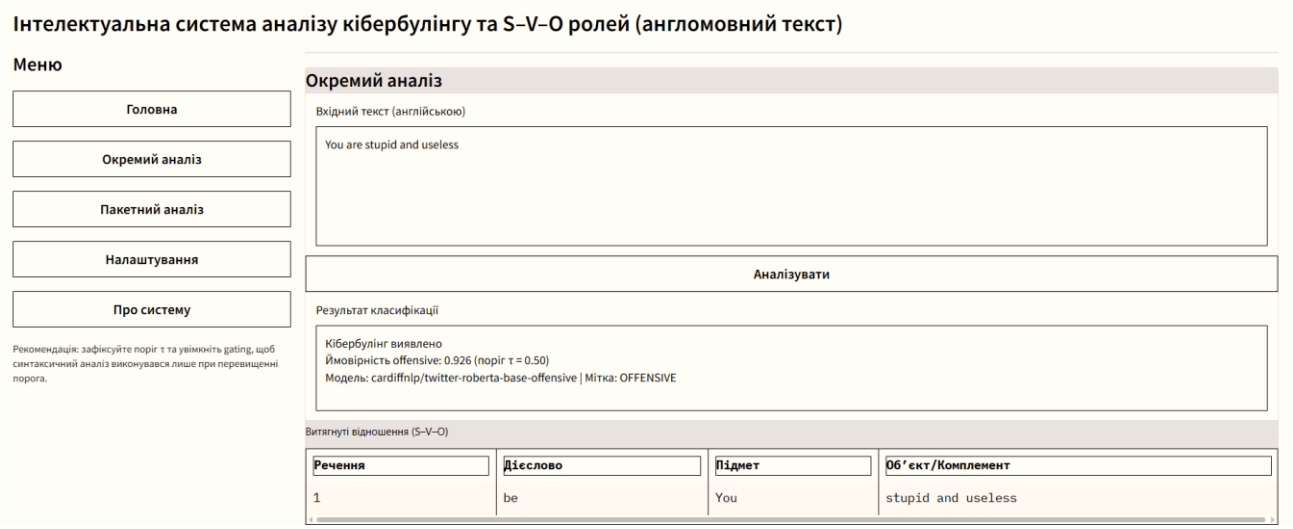


Рисунок 4.6 – Приклад ідентифікації кібербулінгу

Ключовою особливістю експлуатації є керованість параметрів прийняття рішення та пов’язана з цим варіативність поведінки системи у різних сценаріях. За однакового тексту різні значення порога спрацювання можуть приводити до різних висновків, тому користувач повинен розглядати поріг як інструмент налаштування чутливості під політику модерації або аналітичні цілі. Вигляд порогу наведено на рисунку 4.7.

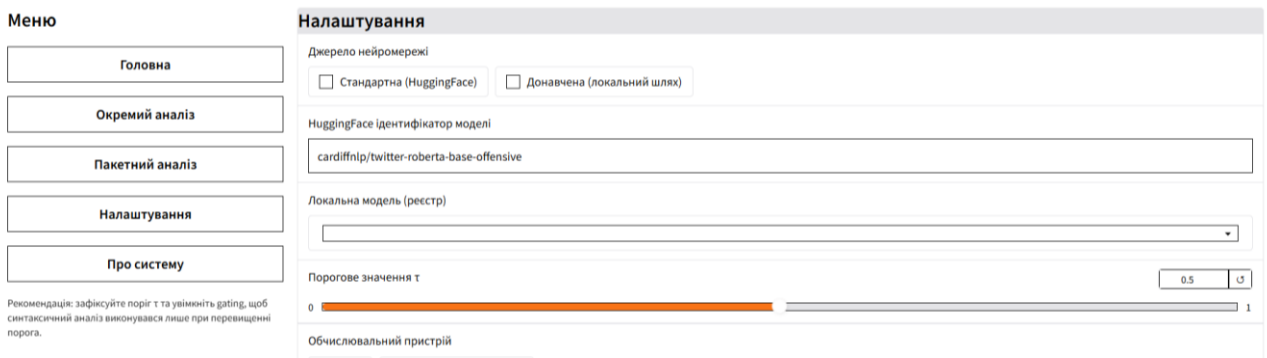


Рисунок 4.7 – Вибір порогового значення

Це проявляється в тому, що для задач пріоритизації перевірки доцільно обирати більш чутливий режим, який зменшує ризик пропуску потенційно проблемного контенту, тоді як для формування офіційних звітів або підготовки вибірок на ручну розмітку доцільним є більш консервативний режим, який знижує частку помилкових спрацювань. Таким чином, користувач взаємодіє не лише з результатом, але й із керуванням компромісом між повнотою виявлення та точністю рішень.

Окремою особливістю є умовність формування пояснювальних результатів. Пояснювальний шар, що містить відношення підмет-дієслово-об’єкт або комплемент, може бути налаштований як такий, що активується лише у випадках, коли система виявляє ознаки кібербулінгу (рисунок 4.8).

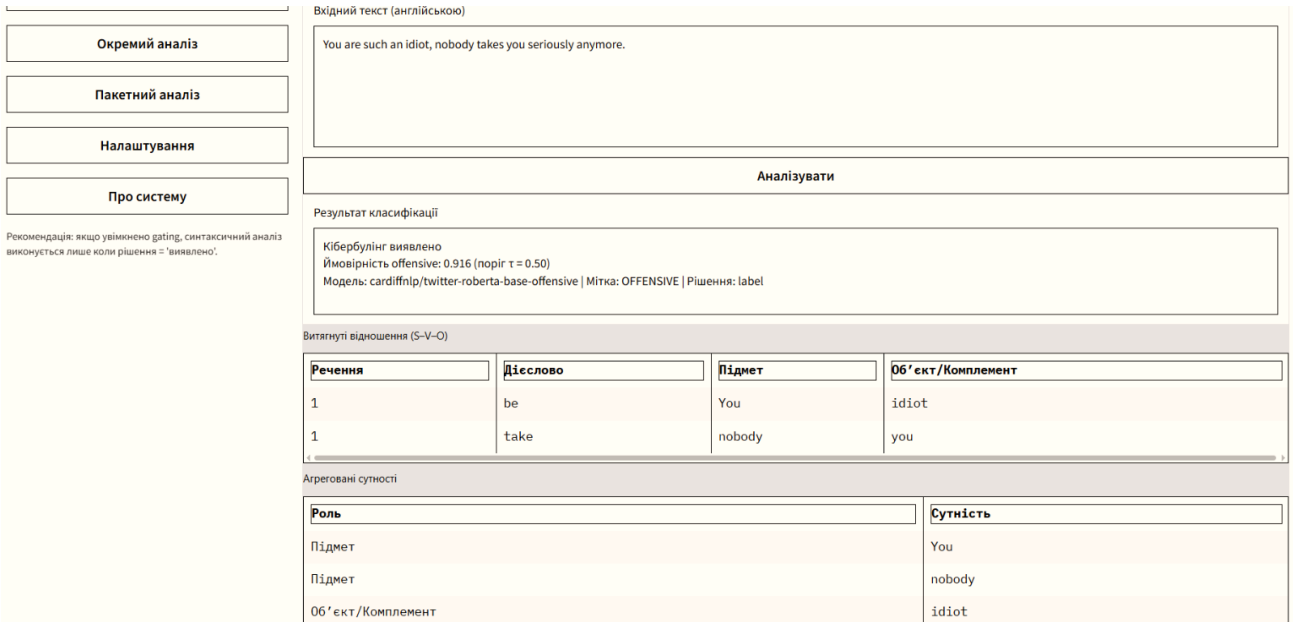


Рисунок 4.8 – Приклад виявлення та аналізу суб’єктів кібербулінгу

Для користувача це означає, що за нейтрального висновку система може не показувати структурні зв’язки, оскільки їх інтерпретація в такому випадку не має практичного сенсу та створює ризик надмірного тлумачення. У результаті користувач отримує більш прогнозовану поведінку системи: пояснення з’являються тоді, коли вони підтримують розбір інциденту, а не тоді, коли вони можуть ввести в оману або перевантажити увагу деталями.

Система має особливість, пов’язану з відмінністю між міткою моделі та числовою оцінкою. У відображенні результатів важливо сприймати мітку як категоріальне рішення моделі та водночас враховувати імовірнісну оцінку як міру впевненості, яка може змінюватися залежно від домену даних, стилістики та структури повідомлення. Для користувача це означає, що близькі до порога значення слід трактувати обережно, особливо у випадках, коли текст містить іронію, цитування, фрагменти діалогу або контекстно залежні натяки. Практично це проявляється як потреба у додатковій перевірці прикордонних випадків та можливість уточнення режиму роботи шляхом корекції порога або використання донавченої моделі, якщо це передбачено процесом експлуатації.

Важливою особливістю користувацького застосування є відмінність між інтерактивним режимом аналізу одного повідомлення та пакетним режимом. Інтерактивний режим орієнтований на швидку перевірку конкретного кейсу, коли користувач потребує оперативного висновку та короткого пояснення для прийняття рішення. Пакетний режим (рисунок 4.9) орієнтований на обробку масивів повідомлень і використовується, коли користувач виконує моніторинг, попереднє сортування або підготовку аналітичної вибірки.

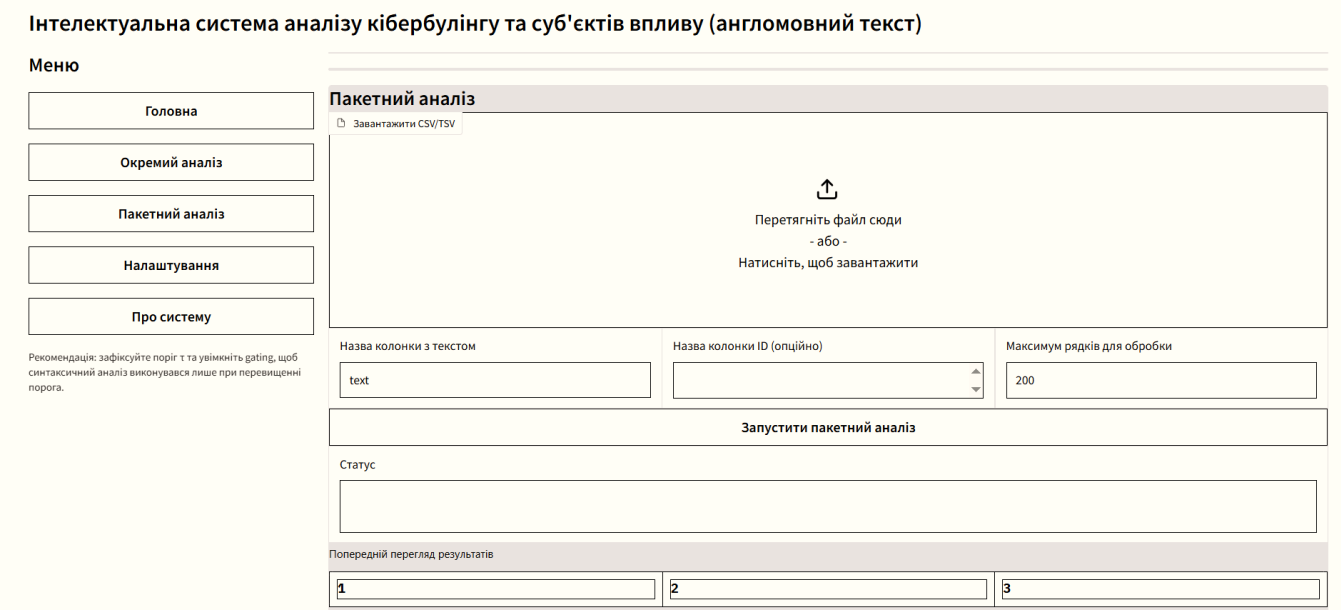


Рисунок 4.9 – Вигляд вікна пакетного режиму

У пакетному режимі результат має розглядатися як структурований вихідний артефакт, придатний для подальшої роботи в табличному середовищі, а не як фінальне рішення щодо кожного повідомлення, оскільки масова обробка підсилює вплив доменних зсувів і може потребувати додаткової верифікації на репрезентативній підвибірці.

Користувач також має враховувати ресурсну специфіку системи та її вплив на час відповіді. Нейромережевий модуль та синтаксичний аналіз створюють різні профілі навантаження, тому при роботі з великими обсягами даних або за обмежених ресурсів доцільним є використання режимів, що зменшують зайві обчислення, зокрема умовне виконання синтаксичного аналізу та обмеження кількості рядків у пакетній обробці. Такий підхід з позиції користувача є не технічною деталлю, а практичною вимогою до стабільної роботи: система повинна залишатися керованою, не зависати на обробці довгих або нестандартних текстів і забезпечувати прогнозований час отримання результату.

Окремою особливістю є те, що пояснювальні структури не гарантують безпомилкового встановлення ролей у кожному повідомленні. Вони відображають результат синтаксичного розбору та його інтерпретації, тому в умовах неформальної мови, порушеної пунктуації, скорочень або фрагментарних висловлювань можливі неповні або спрощені структури. З позиції користувача це означає, що таблиця відношень має інтерпретуватися як допоміжний інструмент для швидкого розуміння потенційної спрямованості впливу, а не як формальний доказ. Практична користь полягає в тому, що навіть неповні структури можуть підсвітити учасників взаємодії та ключові предикати, які варто перевірити в контексті діалогу або модераційного кейсу.

Нарешті, суттєвою особливістю є орієнтація системи на відповідальне використання результатів. Для користувача це означає, що система підтримує прийняття рішення, але не підміняє його, а коректне застосування передбачає співвіднесення автоматизованого висновку з політиками платформи, контекстом комунікації та можливими помилками моделі. Така особливість проявляється у використанні системи для пріоритизації перегляду, підготовки пояснювальних звітів

і первинного сортування контенту, тоді як остаточні дії щодо користувачів або повідомлень мають залишатися в зоні відповідальності людини та визначатися процедурою, прийнятою в конкретному середовищі експлуатації.

4.5 Дослідження ефективності та інтерпретація отриманих результатів

Для дослідження ефективності запропонованого підходу було проведено навчання трансформерної моделі DistilRoBERTa у форматі двокласової класифікації («кібербулінг» / «не кібербулінг»). Як вихідні дані використано об’єднаний корпус, сформований на основі датасетів «Cyberbullying Classification» та «Cyberbullying Detection», попередньо очищений від дублювань, неінформативних і некоректно розмічених прикладів. До тренувальної частини включено 80 % зразків, до валідаційної – 20 %. Навчання виконувалось із використанням оптимізатора AdamW, функції втрат cross-entropy.

Для оцінювання якості моделі застосовано метрики точності (Accuracy), повноти (Recall), точності передбачення (Precision) і збалансованої гармонічної середньої (F_1-score). Нижче наведено результати навчання за трьома фінальними епохами (таблиця 4.1 та 4.2).

Таблиця 4.1 – Динаміка навчання моделі DistilRoBERTa на зведеному корпусі

Епоха	**Train loss**	**Val loss**	**Accuracy**	**Precision**	**Recall**	**F_1**
2	0.312	0.338	0.921	0.914	0.907	0.910
3	0.279	0.324	0.929	0.923	0.916	0.920
4	0.256	0.318	0.934	0.928	0.922	0.925

Динаміка втрат демонструє поступове зниження як тренувальної, так і валідаційної похибки, що свідчить про стійку збіжність без ознак перенавчання. Найбільше покращення спостерігається між другою і третьою епохами, після чого метрики стабілізуються. Значення F_1 = 0.925 підтверджує здатність моделі

ефективно розрізняти повідомлення з ознаками кібербулінгу навіть у випадках стилістичної варіативності та неформального слововжитку.

Порівняльний аналіз показав, що використання корпусу Cyberbullying Classification підвищує специфічність класифікації щодо повідомлень, які містять спрямовану агресію до певних груп, тоді як додавання Cyberbullying Detection покращує загальну чутливість моделі до ширшого спектра образливих висловлювань. Об’єднання цих джерел дозволило отримати збалансований результат, коли модель зберігає високу точність для явних випадків токсичності та водночас не втрачає здатності розпізнавати завуальовані прояви агресії.

Інтерпретація результатів показує, що використання трансформерної архітектури з контекстною токенізацією ефективно з короткими текстами соціальних мереж. Позитивна динаміка F_1-показника у поєднанні з низьким Val loss підтверджує вибір гіперпараметрів і потенціал подальшого донавчання моделі на спеціалізованих доменах (наприклад, україномовних корпусах або аудіотранскриптах), що забезпечить адаптацію системи до інших умов експлуатації.

Подані результати (таблиця 4.2) вказують на те, що підхід зберігає доволі високу якість розпізнавання кібербулінгу за зміни домену вхідних даних.

Таблиця 4.2 – Стійкість до доменного зсуву: якість на підвибірках різного походження ($\tau = 0.50$), кількість зразків – 50.

Джерело повідомлень	**Accuracy**	**Precision (кібербулінг)**	**Recall (кібербулінг)**	**F_1 (кібербулінг)**
Twitter-повідомлення (короткий формат)	0.939	0.934	0.928	0.931
Інші платформи (коментарі, обговорення, повідомлення)	0.926	0.918	0.904	0.911

На підвибірці Twitter-повідомлень отримано Accuracy 0.939 та F_1 0.931, що узгоджується з природою короткого формату, де трансформерні моделі ефективно відпрацьовують контекст навіть за наявності сленгу, скорочень і орфографічних відхилень. Значення Precision 0.934 і Recall 0.928 формують профіль помилок без

різкого зміщення в бік пропусків або надмірних спрацювань, що є важливим для використання у модераційних сценаріях.

Під час переходу до повідомлень з інших платформ метрики зменшуються (Accuracy 0.926, F_1 0.911), що є типовим наслідком доменного зсуву. Коментарі та обговорення частіше містять довші висловлювання, складніші дискурсивні зв’язки, контекстні відсилання та інші стилістичні норми подання агресії, тому частина проявів токсичності має менш прямі лексичні маркери. У такому середовищі Precision 0.918 і Recall 0.904 означають, що певна частка агресивних повідомлень лишається невиявленою, а також з’являються додаткові помилкові спрацювання на контекстно залежних репліках. Водночас F_1 вище 0.90 на вибірці з 50 зразків на домен зберігає практичну придатність підходу, але для стабільнішої роботи в гетерогенних джерелах потрібні доменна адаптація та калібрування порога.

Іншою частиною дослідження стала перевірка виявлення суб'єктів впливу кібербулінгу. Як видно з рисунку 4.10, повідомлення «*You are such an idiot, nobody takes your seriously anymore*», класифіковано як булінг, що є коректно.

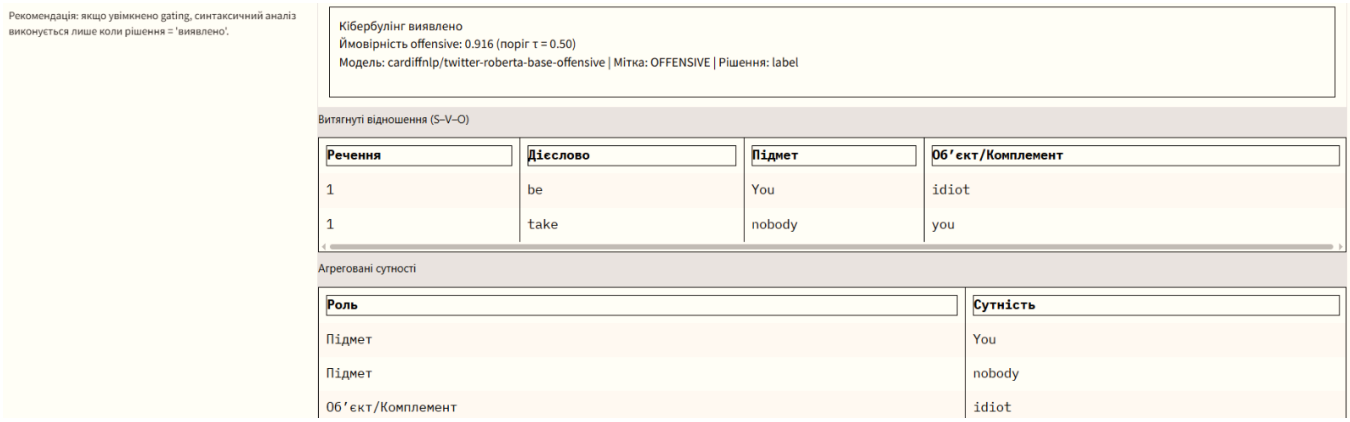


Рисунок 4.10 – Класифікація повідомлення як булінг

Результати синтаксичного аналізу є лінгвістично коректними. Для першого речення правильно ідентифіковано копулятивну конструкцію з дієсловом *be*, де займенник *you* виступає підметом, а іменник *idiot* – іменним комплементом. У другій частині реченні коректно визначено підмет *nobody*, дієслово *take* та об’єкт

you. Таким чином, система показує здатність обробляти як прості, так і складніші синтаксичні структури з узагальненими суб’єктами.

Агрегація сутностей також виконана відповідно до формальних синтаксичних ролей, однак виявляє характерне обмеження суто синтаксичного підходу. Зокрема, лексема *nobody* коректно визначається як граматичний підмет, хоча з точки зору семантики кібербулінгу вона не є реальним суб’єктом впливу, а виконує риторичну функцію підсилення приниження адресата. Дане обмеження не є помилкою реалізації, а відображає загальну проблему інтерпретації імпліцитних та узагальнених акторів у задачах аналізу кібербулінгу.

Загалом отримані результати підтверджують працездатність та внутрішню узгодженість запропонованої системи, а також її придатність для використання як дослідницького прототипу в задачах виявлення кібербулінгу та аналізу суб’єктно-об’єктних відношень у соціальних медіа. Виявлені обмеження окреслюють перспективи подальшого розвитку системи, зокрема шляхом інтеграції семантичних та прагматичних рівнів аналізу.

Висновки до розділу 4

У розділі 4 сформовано цілісне прикладне підґрунтя інтелектуальної системи: від формалізації програмних сутностей і зв’язків між ними до перевірки працездатності реалізації та інтерпретації отриманих експериментальних показників. Запропонована логіка побудована як послідовність, де вхідне повідомлення переходить у структуровані представлення, результати класифікації й пояснювальні властивості, які повертаються користувачеві в уніфікованому форматі.

Наведено програмну структуру системи на рівні моделі даних і середовища виконання. Діаграма класів визначає, які сутності зберігаються під час обробки, як вони агрегуються у результат запиту та як організовано життєвий цикл похідних об’єктів. Архітектурна схема середовища доповнює модель даних розподілом відповідальностей між інтерфейсом, координатором обробки, модулем

нейромережевої класифікації, синтаксичним аналізом і табличною обробкою, а також уточнює зовнішні залежності від сховища даних та джерела моделей. У сукупності це задає відтворювану структуру запуску в середовищі Colab і визначає межі компонентів, що впливають на стабільність роботи.

Наведено прикладні алгоритми, які формалізують критичні процедури реалізації, що безпосередньо визначають коректність результатів у користувацькому сценарії. Окремо виділено ініціалізацію середовища, правила вибору моделі, реєстрацію локальних артефактів, нормалізацію виходу класифікатора та кероване ухвалення рішення за параметрами. Таке виділення процедур зменшує ризик логічних розбіжностей під час модифікації коду, а також дозволяє локалізувати джерела помилок на рівні чітко визначених етапів конвеєра. Умовне виконання інтерпретаційного шару використано як практичний механізм керування обчислювальними витратами та уникнення надмірної деталізації у випадках, коли класифікаційне рішення не потребує пояснення.

Реалізовано прикладне тестування у формі модульних перевірок, спрямованих на надійність експлуатаційних сценаріїв. Перевірено коректність перетворення оцінок і міток класифікатора, валідацію налаштувань для різних джерел моделі, обробку порожнього вводу, поведінку інтерпретаційного шару за негативного рішення, формування табличних виходів і створення файлів експорту в пакетному режимі. Успішне проходження тестів означає, що система передбачувано реагує на типові помилки конфігурації, формує узгоджені вихідні структури для інтерфейсу.

Описано особливості використання системи з позиції користувача, де основний акцент зроблено на правильному читанні результатів і керуванні параметрами. Користувацький сценарій включає інтерпретацію імовірнісної оцінки разом із категоріальним рішенням, контроль чутливості через поріг, розуміння умов формування пояснювальних структур і вибір між інтерактивною перевіркою одиничного повідомлення та пакетною обробкою масивів даних. Така постановка підкреслює, що система працює як інструмент підтримки рішення, а практична

корисність зростає за коректного налаштування параметрів під політику модерації або аналітичну мету, з урахуванням ресурсних обмежень середовища виконання.

Наведено узагальнену оцінку ефективності на основі навчання трансформерної моделі у задачі двокласової класифікації та аналізу стійкості за зміни джерела повідомлень. Динаміка навчання за епохами характеризується спаданням втрат і приростом показників якості, що відповідає очікуваному покращенню параметрів моделі в межах обмеженого числа ітерацій. Окремо оцінено якість на підвибірках різного походження: для коротких повідомлень отримано вищі значення метрик, тоді як для текстів з інших платформ спостерігається зниження, пов’язане з відмінностями стилю, довжини й контекстної насиченості повідомлень. Розбір прикладу з інтерпретаційним виходом виявляє типове обмеження суто синтаксичної атрибуції учасників взаємодії, коли граматично коректні ролі не завжди відповідають прагматичній природі агресії; цей ефект слід враховувати під час практичного застосування.

Загалом, у розділі отримано прикладну реалізацію інтелектуальної системи, у якій формалізовано структуру даних, описано алгоритмічний склад компонентів, перевірено ключові експлуатаційні сценарії та наведено показники якості, достатні для обґрунтування працездатності прототипу. Виявлені особливості використання та обмеження інтерпретації визначають напрям подальшого розвитку: адаптацію під конкретні платформи через донавчання, калібрування порогів і розширення інтерпретаційного шару за рахунок семантичних та прагматичних ознак, що підвищить точність атрибуції учасників взаємодії у реальних комунікативних контекстах.

Загальні висновки

Метою кваліфікаційної роботи було підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб’єктів впливу та спрямованості взаємодії на основі трансформерних моделей. Поставлену мету було досягнуто.

Для досягнення мети були вирішити такі задачі:

– досліджено сучасний стан області виявлення кібербулінгу;

– виконано огляд сучасних методів та засобів виявлення кібербулінгу та суб’єктів впливу;

– виконано аналіз наукових досліджень предметної області;

– розроблено метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконано підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу;

– здійснено програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;

– виконано дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

Проведено дослідження ефективності запропонованого підходу та інтерпретацію результатів на підвибірках різного походження. Експериментальна частина включала оцінювання за стандартними метриками класифікації та аналіз чутливості до зміни домену вхідних повідомлень. Окремо проаналізовано інтерпретаційні виходи на прикладах, що дозволило окреслити межі застосування синтаксично орієнтованого пояснення у випадках узагальнених або контекстно залежних акторів. Отримані результати підтверджують придатність запропонованої реалізації як дослідницького прототипу для задач виявлення кібербулінгу з формуванням пояснювального подання взаємодії учасників.

Обмеження підходу пов’язані з доменом і форматом даних та природою інтерпретації: система найкраще працює на англомовних коротких повідомленнях,

тоді як за переходу на інші платформи, інші мовні стилі або довші тексти можливе просідання якості й потреба у калібруванні порога та/або донавчанні під конкретний домен; пояснювальний шар ґрунтується на синтаксичному розборі, тому в неформальних репліках із помилками, скороченнями, іронією чи цитуванням інколи формуються неповні або прагматично неоднозначні зв’язки між учасниками, а мультимодальні сигнали (емодзі як носії прагматики, зображення, GIF) не враховуються; у пакетному режимі додатково діють ресурсні обмеження, тому для великих обсягів даних потрібні ліміти та контроль часу виконання.

Практичне значення роботи полягає у створенні програмного інструмента, який надає користувачеві не лише факт виявлення кібербулінгу, а й структуровані пояснювальні результати, придатні для пріоритизації перегляду контенту, формування аналітичних вибірок і підготовки звітів. Подальший розвиток підходу доцільно пов’язати з адаптацією до конкретних платформ через донавчання на доменних корпусах, калібруванням порогових параметрів під задані політики ризику та розширенням інтерпретаційного шару за рахунок семантичних і прагматичних ознак, що дозволить точніше враховувати контекст, іронію та непрямі мовні акти в реальній онлайн-комунікації.

За темою кваліфікаційної роботи підготовлено до публікації статтю в фаховому виданні категорії Б. Основні наукові й практичні результати роботи доповідались у доповіді «Трансформерне виявлення суб’єктів кібербулінгу за текстовими повідомленнями» [70] на XVII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп’ютерних наук АПКН-2025» (м. Хмельницький) 14-15 листопада 2025 року.

Перелік посилань

1. Giumetti G. W., Kowalski R. M. Cyberbullying via social media and well-being. *Current Opinion in Psychology*. 2022. Vol. 45. P. 101314. URL: https://doi.org/10.1016/j.copsyc.2022.101314 (date of access: 14.10.2025).

2. Cyberbullying and mental health: past, present and future / S. Bansal et al. Frontiers in Psychology. 2024. Vol. 14. URL: https://doi.org/10.3389/fpsyg.2023.1279234 (date of access: 14.10.2025).

3. *GAY med i kampen mot mobbning | Friends*. URL: https://friends.se/uploads/sites/2/2024/08/Bullying-An-Inclusive-definition-UNESCO-WABF.pdf (дата звернення: 14.10.2025).

4. Kee D. M. H., Al- Anesi M. A. L., Al- Anesi S. A. L. Cyberbullying on social media under the influence of COVID- 19. *Global Business and Organizational Excellence*. 2022. URL: https://doi.org/10.1002/joe.22175 (date of access: 14.10.2025).

5. Giumetti G. W., Kowalski R. M. Cyberbullying via social media and well-being. *Current Opinion in Psychology*. 2022. Vol. 45. P. 101314. URL: https://doi.org/10.1016/j.copsyc.2022.101314 (date of access: 14.10.2025).

6. The EU's Digital Services Act. *European Commission*. URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (date of access: 14.10.2025).

7. Yemima C. K. The Forms of Cyberbullying Behavior among Teenage Students: A Systematic Literature Review. *Jurnal Bimbingan dan Konseling Terapan*. 2023. Vol. 7, no. 2. P. 151. URL: https://doi.org/10.30598/jbkt.v7i2.1745 (date of access: 14.10.2025).

8. Arisanty M., Wiradharma G. The motivation of flaming perpetrators as cyberbullying behavior in social media. *Jurnal Kajian Komunikasi*. 2022. Vol. 10, no. 2. P. 215. URL: https://doi.org/10.24198/jkk.v10i2.39876 (date of access: 14.10.2025).

9. Gong J., Yang L. A Review on Flaming Ignition of Solid Combustibles: Pyrolysis Kinetics, Experimental Methods and Modelling. *Fire Technology*. 2022. URL: https://doi.org/10.1007/s10694-022-01339-7 (date of access: 14.10.2025).

10. Kizza J. M. Cyberbullying, Cyberstalking and Cyber Harassment. *Undergraduate Topics in Computer Science*. Cham, 2023. P. 199–210. URL: https://doi.org/10.1007/978-3-031-31906-8_9 (date of access: 14.10.2025).

11. Milosevic T., Van Royen K., Davis B. Artificial Intelligence to Address Cyberbullying, Harassment and Abuse: New Directions in the Midst of Complexity. *International Journal of Bullying Prevention*. 2022. Vol. 4, no. 1. P. 1–5. URL: https://doi.org/10.1007/s42380-022-00117-x (date of access: 14.10.2025).

12. Slutskiy P. Defamation, Libel and Slander. *Communication and Libertarianism*. Singapore, 2021. P. 337–350. URL: https://doi.org/10.1007/978-981-33-6664-0_22 (date of access: 14.10.2025).

13. Al-Zoubi M. Crimes of Electronic Defamation, Libel, and Slander under Jordanian Cybercrimes Law. *International Review of Law*. 2023. Vol. 12, no. 1. P. 267–284. URL: https://doi.org/10.29117/irl.2023.0260 (date of access: 14.10.2025).

14. Categorization of Cyberbullying based on Intentional Dimension / M. F. Gan et al. *2024 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Shah Alam, Malaysia, 29 June 2024. 2024. P. 285–290. URL: https://doi.org/10.1109/i2cacis61270.2024.10649619 (date of access: 14.10.2025).

15. Saragih D. C., Windarwati H. D., Merdikawati A. Are Personality Types Related to Cyberbullying Behavior Trends in Adolescents?. *Jurnal Keperawatan Jiwa*. 2020. Vol. 8, no. 3. P. 307. URL: https://doi.org/10.26714/jkj.8.3.2020.307-318 (date of access: 14.10.2025).

16. Cyberbullying and Cyberviolence Detection: A Triangular User-Activity-Content View / S. Wang et al. *IEEE/CAA Journal of Automatica Sinica*. 2022. Vol. 9, no. 8. P. 1384–1405. URL: https://doi.org/10.1109/jas.2022.105740 (date of access: 14.10.2025).

17. Cyber-Ostracism, Depression, and Adolescents' Cyberbullying Perpetration: A Cross-Lagged Panel Analysis / H. Ding et al. *Youth & Society*. 2024. URL: https://doi.org/10.1177/0044118x241301062 (date of access: 14.10.2025).

18. Özişli Ö. The Effect Of Workplace Ostracism On Organizational Silence And Workplace Loneliness A Study On Healthcare Workers. *International Journal of Health*

Services Research and Policy. 2022. URL: https://doi.org/10.33457/ijhsrp.1131522 (date of access: 14.10.2025).

19. Stalking and cyberstalking among college students: Prevalence and distinctions between harassment, victimization, and perpetration. / R. C. Garthe et al. *Psychology of Violence*. 2025. URL: https://doi.org/10.1037/vio0000629 (date of access: 14.10.2025).

20. Factors Influencing the Occurrence of Cyberbullying on Facebook among Undergraduate Students in Kenyan Universities / E. Odhiambo Ogolla et al. *Issue 6*. 2022. Vol. 3, no. 6. P. 109–120. URL: https://doi.org/10.46606/eajess2022v03i06.0242 (date of access: 14.10.2025).

21. Children's Online Safety: Predictive Factors of Cyberbullying and Online Grooming Involvement / A. Tintori et al. *Societies*. 2023. Vol. 13, no. 2. P. 47. URL: https://doi.org/10.3390/soc13020047 (date of access: 14.10.2025).

22. Cyberbullying Characteristics and Prevention–What Can We Learn from Narratives Provided by Adolescents and Their Teachers? / J. Pyżalski et al. *International Journal of Environmental Research and Public Health*. 2022. Vol. 19, no. 18. P. 11589. URL: https://doi.org/10.3390/ijerph191811589 (date of access: 14.10.2025).

23. Characteristics and effectiveness of interventions to reduce cyberbullying: a systematic review / J. Henares-Montiel et al. *Frontiers in Public Health*. 2023. Vol. 11. URL: https://doi.org/10.3389/fpubh.2023.1219727 (date of access: 14.10.2025).

24. Yi P., Zubiaga A. Session-Based Cyberbullying Detection in Social Media: A Survey. *SSRN Electronic Journal*. 2022. URL: https://doi.org/10.2139/ssrn.4208013 (date of access: 14.10.2025).

25. Ray G., McDermott C. D., Nicho M. Cyberbullying on Social Media: Definitions, Prevalence, and Impact Challenges. *Journal of Cybersecurity*. 2024. Vol. 10, no. 1. URL: https://doi.org/10.1093/cybsec/tyae026 (date of access: 14.10.2025).

26. Zhao Y., Chu X., Rong K. Cyberbullying experience and bystander behavior in cyberbullying incidents: The serial mediating roles of perceived incident severity and empathy. *Computers in Human Behavior*. 2022. P. 107484. URL: https://doi.org/10.1016/j.chb.2022.107484 (date of access: 14.10.2025).

27. Bullying and Cyberbullying in School: Rapid Review on the Roles of Gratitude, Forgiveness, and Self-Regulation / W. A. d. Oliveira et al. *International Journal of Environmental Research and Public Health*. 2024. Vol. 21, no. 7. P. 839. URL: https://doi.org/10.3390/ijerph21070839 (date of access: 14.10.2025).

28. Polanco-Levicán K., Salvo-Garrido S. Bystander Roles in Cyberbullying: A Mini-Review of Who, How Many, and Why. *Frontiers in Psychology*. 2021. Vol. 12. URL: https://doi.org/10.3389/fpsyg.2021.676787 (date of access: 14.10.2025).

29. Bias and Cyberbullying Detection and Data Generation Using Transformer Artificial Intelligence Models and Top Large Language Models / Y. Kumar et al. *Electronics*. 2024. Vol. 13, no. 17. P. 3431. URL: https://doi.org/10.3390/electronics13173431 (date of access: 14.10.2025).

30. OffensEval Shared Task - OLID. *Google Drive: Sign-in*. URL: https://sites.google.com/site/offensevalsharedtask/olid (date of access: 14.10.2025).

31. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) / M. Zampieri et al. *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/s19-2010 (date of access: 14.10.2025).

32. SemEval-2021 Task 5: Toxic Spans Detection / J. Pavlopoulos et al. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Online. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.semeval-1.6 (date of access: 14.10.2025).

33. Jacobs G., Van Hee C., Hoste V. Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?. *Natural Language Engineering*. 2020. P. 1–26. URL: https://doi.org/10.1017/s135132492000056x (date of access: 14.10.2025).

34. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation / D. Ghosal et al. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/d19-1015 (date of access: 14.10.2025).

35. Wang L., Zhang L. Hawkes processes for understanding heterogeneity in information propagation on Twitter. *Frontiers in Physics*. 2022. Vol. 10. URL: https://doi.org/10.3389/fphy.2022.1019380 (date of access: 14.10.2025).

36. Council of Europe – Online resources. *Conseil de l'Europe – Ressources en lignes*. URL: https://edoc.coe.int/en (date of access: 14.10.2025).

37. Regulation - 2016/679 - EN - gdpr - EUR-Lex. *EUR-Lex – Access to European Union law – choose your language*. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (date of access: 14.10.2025).

38. The Digital Services Act package. *Shaping Europe's digital future*. URL: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (date of access: 14.10.2025).

39. *UNESCO Science Report: the race against time for smarter development; executive summary. UNESCO*. URL: https://unesdoc.unesco.org/ark:/48223/pf0000377250 (date of access: 14.10.2025).

40. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification / D. Borkan et al. *WWW '19: The Web Conference*, San Francisco USA. New York, NY, USA, 2019. URL: https://doi.org/10.1145/3308560.3317593 (date of access: 14.10.2025).

41. Ethics guidelines for trustworthy AI. *Shaping Europe's digital future*. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (date of access: 14.10.2025).

42. Про захист персональних даних. *Офіційний вебпортал парламенту України*. URL: https://zakon.rada.gov.ua/laws/show/2297-17 (дата звернення: 14.10.2025).

43. Про освіту. *Офіційний вебпортал парламенту України*. URL: https://zakon.rada.gov.ua/laws/show/2145-19 (дата звернення: 14.10.2025).

44. Про інформацію. *Офіційний вебпортал парламенту України*. URL: https://zakon.rada.gov.ua/laws/show/2657-12 (дата звернення: 14.10.2025).

45. OffensEval 2023: Offensive language identification in the age of Large Language Models / M. Zampieri et al. *Natural Language Engineering*. 2023. Vol. 29, no. 6. P. 1416–1435. URL: https://doi.org/10.1017/s1351324923000517 (date of access: 14.10.2025).

46. Luu S. T., Nguyen N. UIT-ISE-NLP at SemEval-2021 Task 5: Toxic Spans Detection with BiLSTM-CRF and ToxicBERT Comment Classification. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Online. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.semeval-1.113 (date of access: 14.10.2025).

47. Sihab-Us-Sakib S., Rahman M. R., Forhad M. S. A., Aziz M. A. Cyberbullying detection of resource constrained language from social media using transformer-based approach. *Natural Language Processing Journal*, 2024, Vol. 9, 100104. URL: https://doi.org/10.1016/j.nlp.2024.100104 (date of access: 14.10.2025).Aliyeva Ç. O., Yağanoğlu M. Deep learning approach to detect cyberbullying on twitter. *Multimedia Tools and Applications*. 2024. URL: https://doi.org/10.1007/s11042-024-19869-3 (date of access: 14.10.2025).

48. Gutiérrez-Batista K., Gómez-Sánchez J., Fernandez-Basso C. Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model. *Social Network Analysis and Mining*. 2024. Vol. 14, no. 1. URL: https://doi.org/10.1007/s13278-024-01291-0 (date of access: 14.10.2025).

49. Yi P., Zubiaga A. Session-Based Cyberbullying Detection in Social Media: A Survey. *SSRN Electronic Journal*. 2022. URL: https://doi.org/10.2139/ssrn.4208013 (date of access: 14.10.2025).

50. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection / B. Mathew et al. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021. Vol. 35, no. 17. P. 14867–14875. URL: https://doi.org/10.1609/aaai.v35i17.17745 (date of access: 14.10.2025).

51. Hate-speech-CNERG/hatexplain · Datasets at Hugging Face. *Hugging Face – The AI community building the future.* URL: https://huggingface.co/datasets/Hate-speech-CNERG/hatexplain (date of access: 14.10.2025).

52. Site Agnostic Approach to Early Detection of Cyberbullying on Social Media Networks / M. López-Vizcaíno et al. *Sensors*. 2023. Vol. 23, no. 10. P. 4788. URL: https://doi.org/10.3390/s23104788 (date of access: 14.10.2025).

53. Cyberbullying Classification. *Kaggle*. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

54. (date of access: 14.10.2025).

55. Cyberbullying Detection. *Kaggle*. URL: https://www.kaggle.com/datasets/gbiamgaurav/cyberbullying-detection

56. (date of access: 14.10.2025).

57. Kashyap P. Understanding Precision, Recall, and F1 Score Metrics. *Medium*. URL: https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093 (date of access: 14.10.2025).

58. Precision vs. Recall in Machine Learning: Whats the Difference?. *Coursera*. URL: https://www.coursera.org/articles/precision-vs-recall-machine-learning (date of access: 14.10.2025).

59. Ultralytics. F1-Score: Definition, Formula & Applications | Ultralytics. *Ultralytics | Revolutionizing the World of Vision AI*. URL: https://www.ultralytics.com/glossary/f1-score (date of access: 14.10.2025).

60. Tigerschiold T. What is Accuracy, Precision, Recall and F1 Score?. *Elevate your Customer Interactions | Labelf AI*. URL: https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score (date of access: 14.10.2025).

61. What is a Confusion Matrix? | Machine Learning Glossary | Encord | Encord. *Encord | Manage, Curate, and Annotate Data for Multimodal AI*. URL: https://encord.com/glossary/confusion-matrix/ (date of access: 14.10.2025).

62. What is AUC-ROC. *H2O.ai | Convergence of the World's Best Predictive and Generative AI for Private, Protected Data*. URL: https://h2o.ai/wiki/auc-roc/ (date of access: 14.10.2025).

63. Colab.google. *colab.google*. URL: https://colab.google/ (date of access: 14.10.2025).

64. Transformers. *Hugging Face – The AI community building the future*. URL: https://huggingface.co/docs/transformers/index (date of access: 14.10.2025).

65. pandas - Python Data Analysis Library. *pandas - Python Data Analysis Library*. URL: https://pandas.pydata.org/ (date of access: 14.10.2025).

66. typing Support for type hints. *Python documentation*. URL: https://docs.python.org/3/library/typing.html (date of access: 14.10.2025).

67. Gradio. *Gradio*. URL: https://gradio.app/ (date of access: 14.10.2025).

68. GitHub - stanfordnlp/stanza: Stanford NLP Python library for tokenization, sentence segmentation, NER, and parsing of many human languages. *GitHub*. URL: https://github.com/stanfordnlp/stanza (date of access: 14.10.2025).

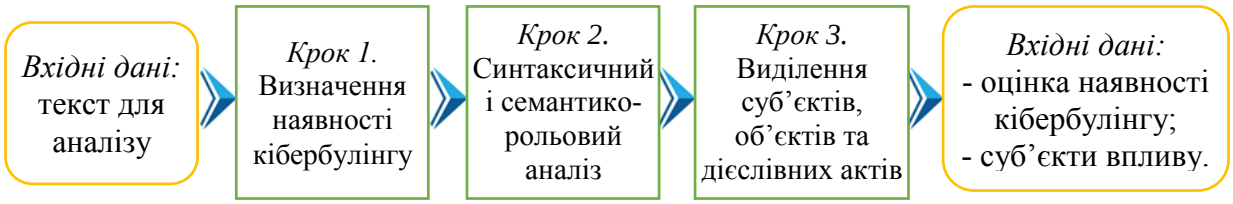
69. Nordquist R. What is Basic Word Order in English?. *ThoughtCo*. URL: https://www.thoughtco.com/subject-verb-object-1692011 (date of access: 14.10.2025).

70. Трансформерне виявлення суб’єктів кібербулінгу за текстовими повідомленнями/ Андрощук В.І., Молчанова М.О. // *Актуальні проблеми комп’ютерних наук : зб. наук. пр. за матеріалами XVII Всеукр. наук.-практ. конф. (АПКН-2025).* – Хмельницький, 14–15 листоп. 2025 р. – Хмельницький, 2025. – С. 15–19.

ДОДАТКИ

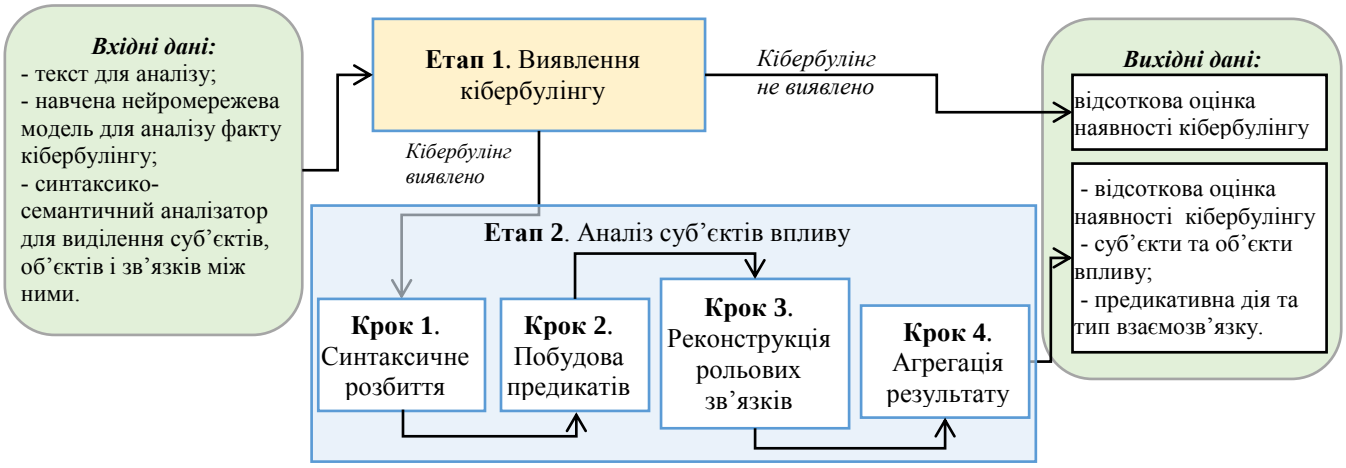
Додаток А

Схема та кроки підходу до багаторівневого виявлення суб'єктів впливу кібербулінгу



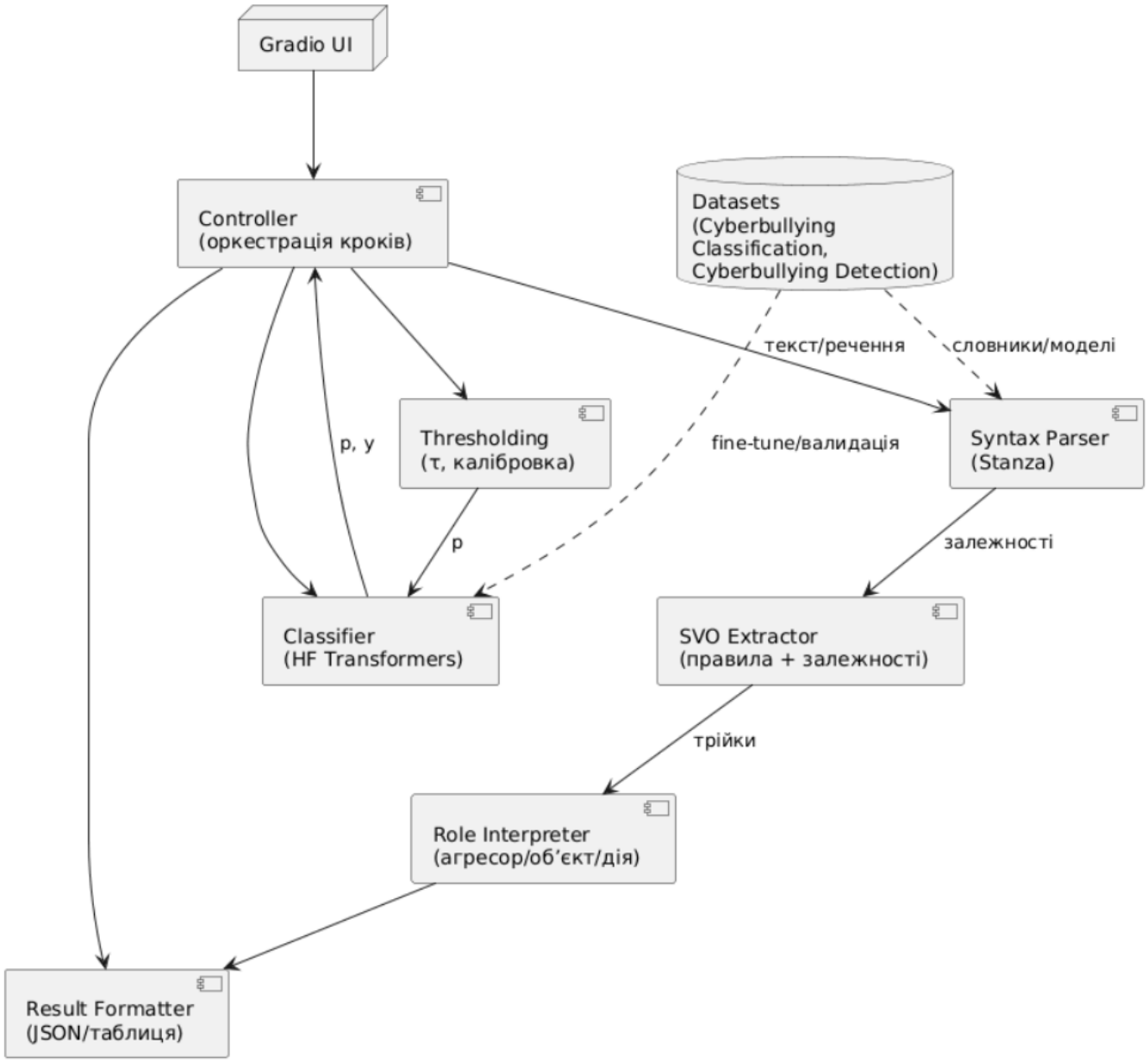
Додаток Б

Схема методу багаторівневого виявлення суб'єктів впливу кібербулінгу



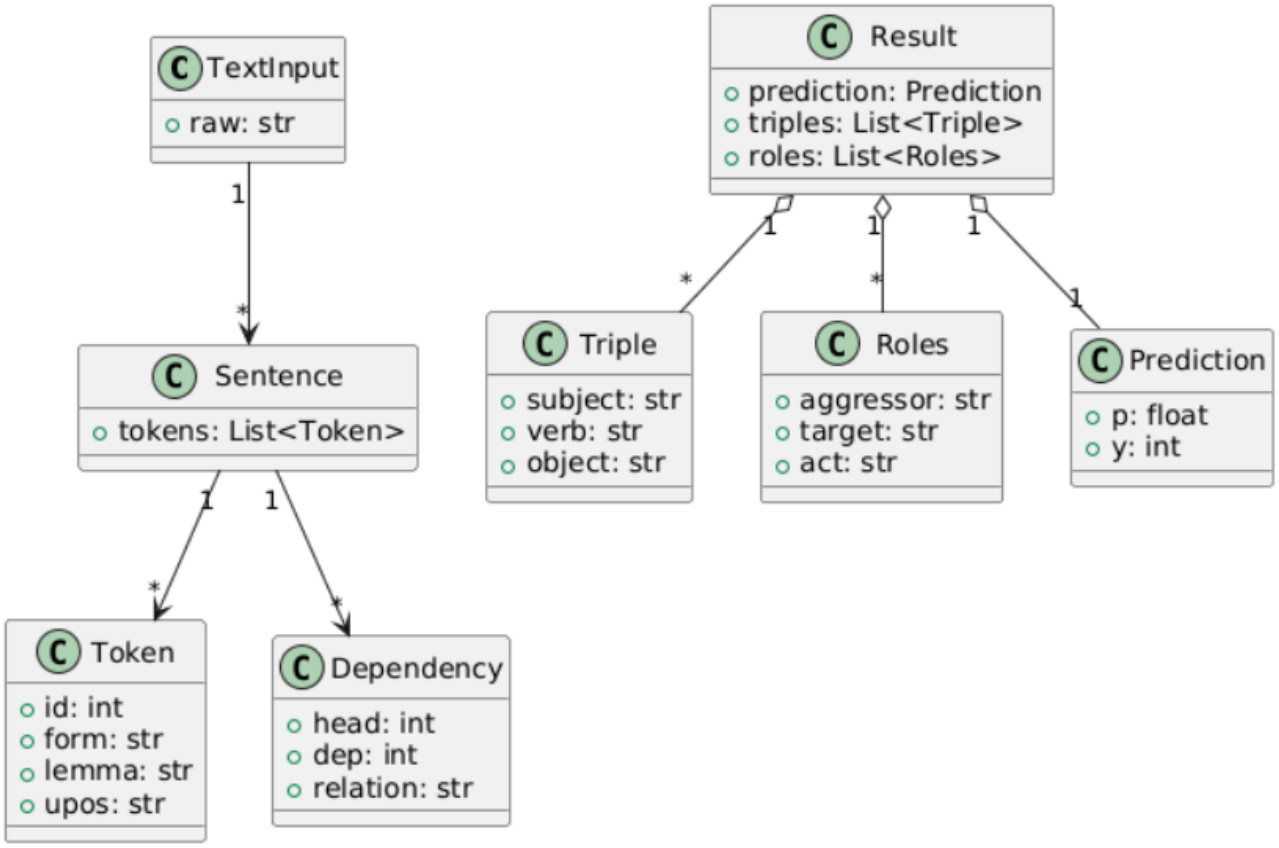
Додаток В

Схема взаємодії компонентів інтелектуальної системи



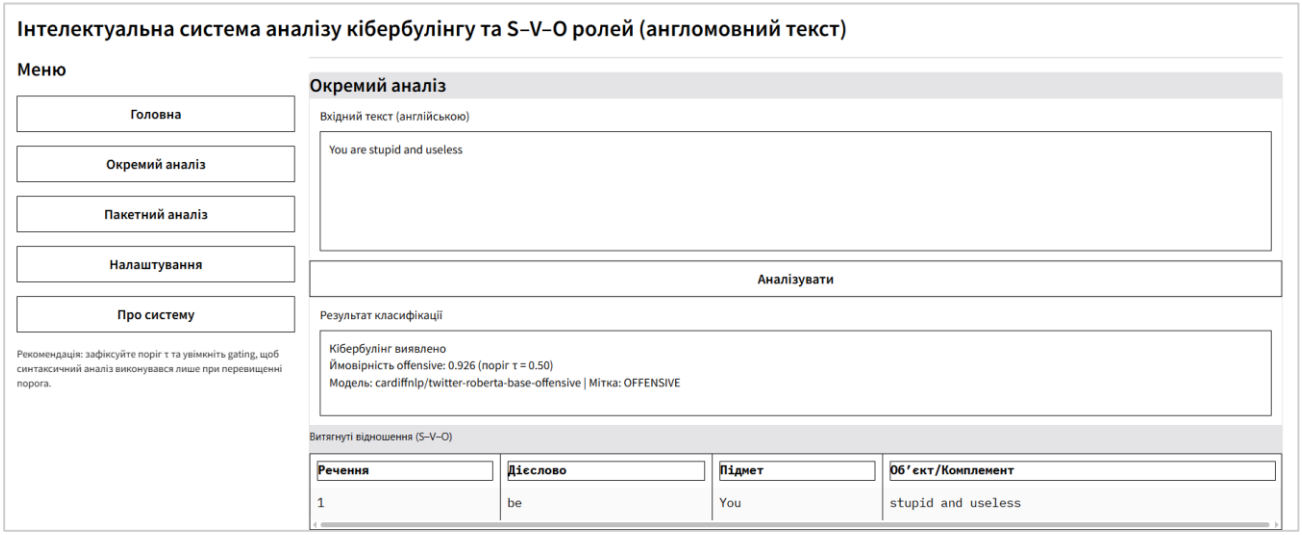
Додаток Г

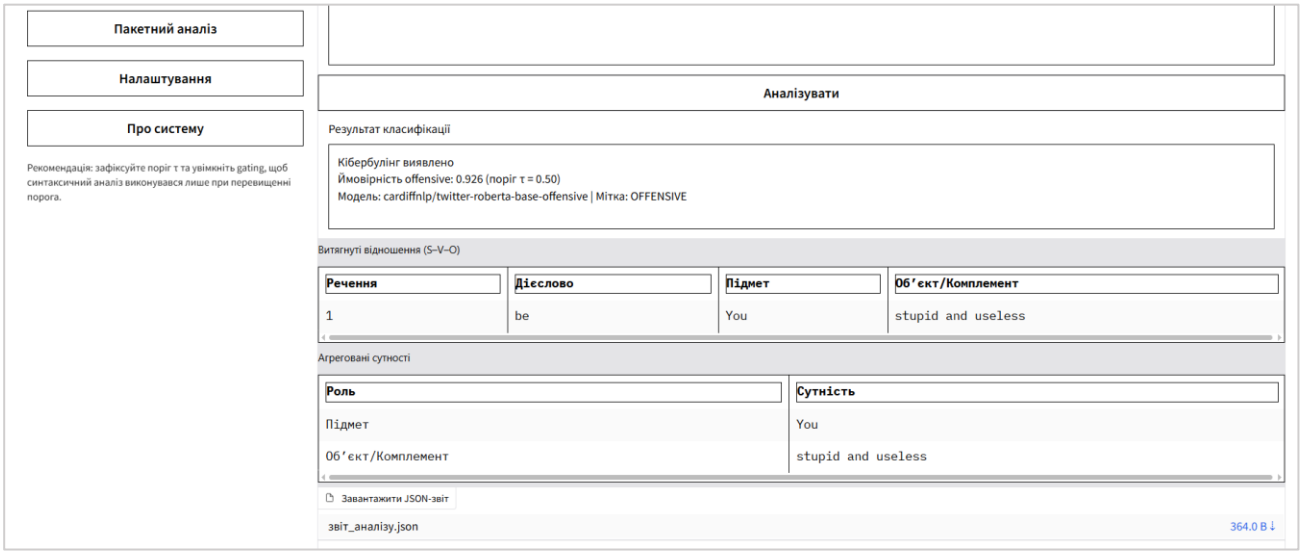
Діаграма класів інтелектуальної системи

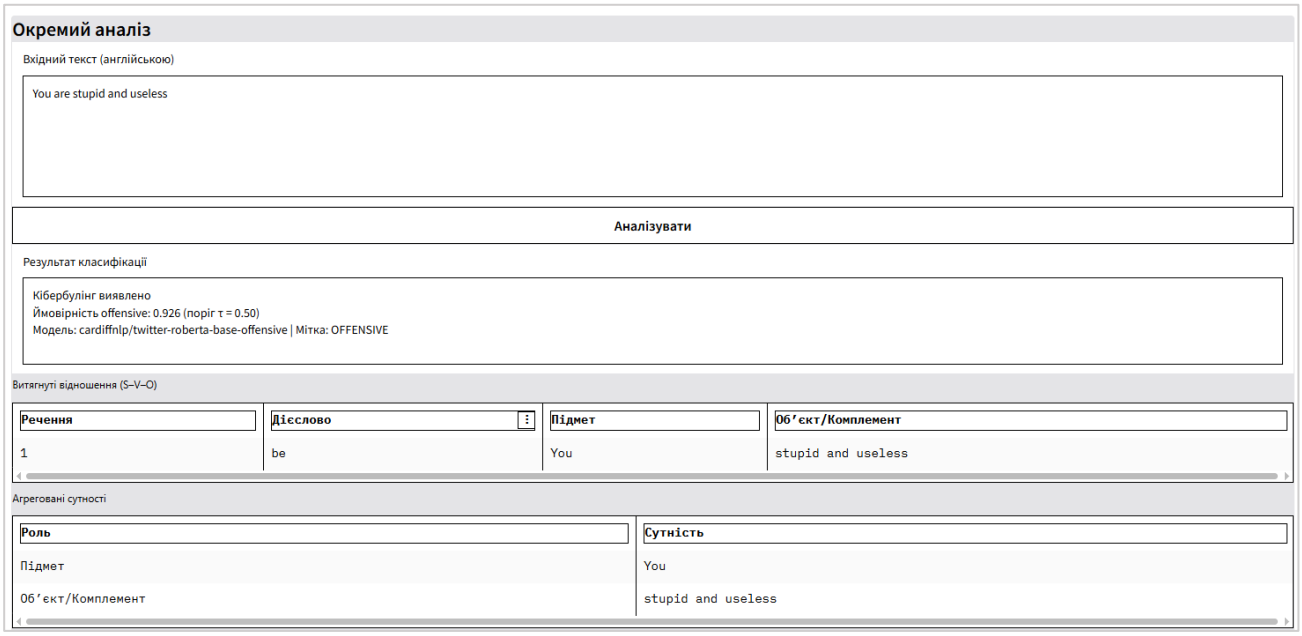


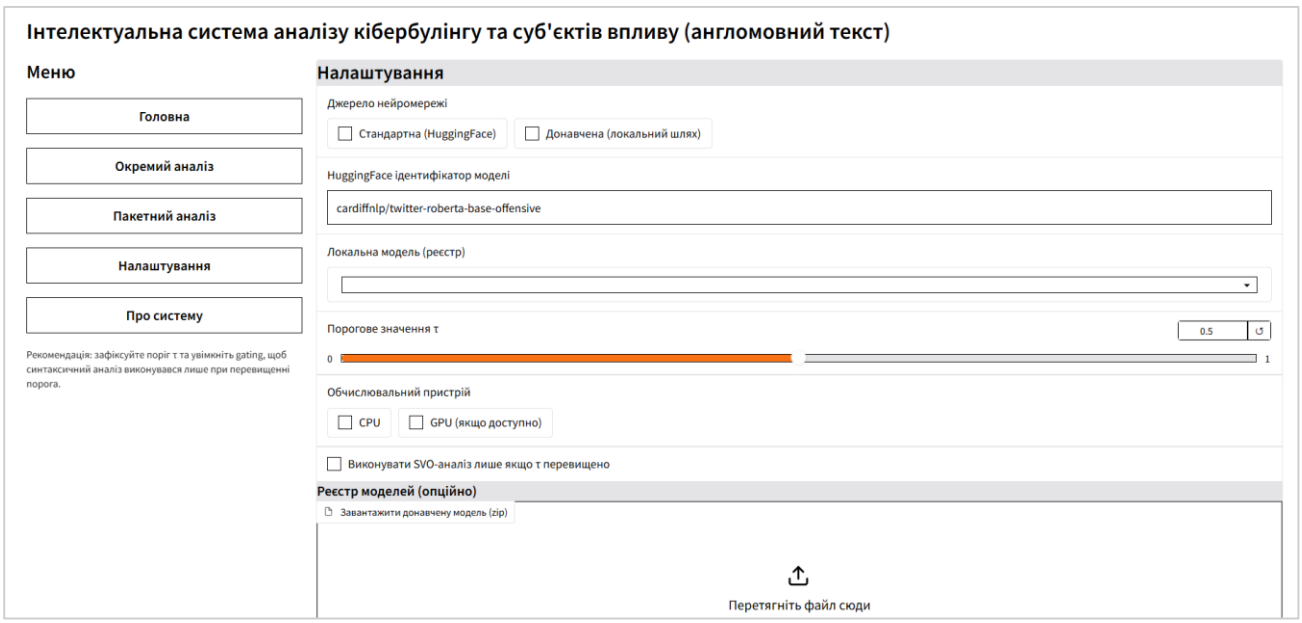
Додаток Д

Світлини екрану інтелектуальної системи






Окремий аналіз
Вхідний текст (англійською)
You are stupid and useless
Аналізувати
Результат класифікації
Кібербулінг виявлено
Ймовірність offensive: 0.926 (поріг τ = 0.50)
Модель: cardiffnlp/twitter-roberta-base-offensive | Мітка: OFFENSIVE
Витягнуті відношення (S–V–O)
Речення
Дієслово
Підмет
Об'єкт/Комплемент
1
be
You
stupid and useless
Агреговані сутності
Роль
Сутність
Підмет
You
Об'єкт/Комплемент
stupid and useless


Інтелектуальна система аналізу кібербулінгу та суб'єктів впливу (англомовний текст)
Меню
Головна
Окремий аналіз
Пакетний аналіз
Налаштування
Про систему
Рекомендація: зафіксуйте поріг τ та увімкніть gating, щоб синтаксичний аналіз виконувався лише при перевищенні порога.
Налаштування
Джерело нейромережі
Стандартна (HuggingFace)
Донавчена (локальний шлях)
HuggingFace ідентифікатор моделі
cardiffnlp/twitter-roberta-base-offensive
Локальна модель (реєстр)
Порогове значення τ
0.5
0
1
Обчислювальний пристрій
CPU
GPU (якщо доступно)
Виконувати SVO-аналіз лише якщо τ перевищено
Реєстр моделей (опційно)
Завантажити донавчену модель (zip)
Перетягніть файл сюди

Додаток Е

Програмні коди

Код інтелектуальної системи багаторівневого виявлення суб'єктів впливу кібербулінгу доступний у репозиторії GitHub: https://github.com/vlad-androshchuk/bullying (дата звернення: 13.12.2025).

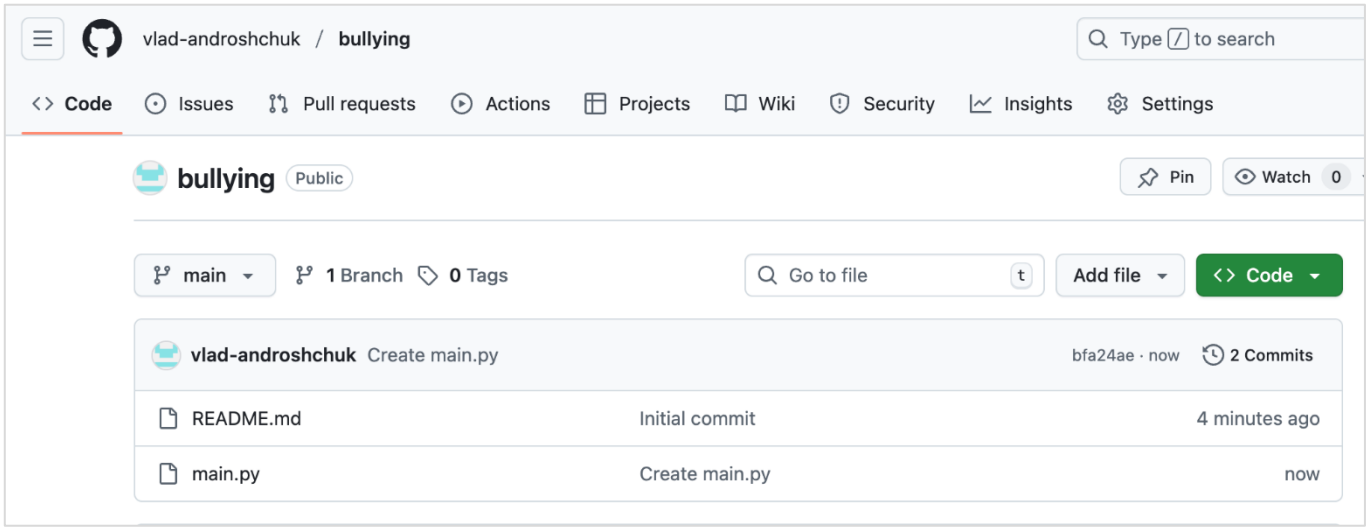


Рисунок Е.1 – Сторінка репозиторію автора

Додаток Ж

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

1. Трансформерне виявлення суб’єктів кібербулінгу за текстовими повідомленнями/ Андрощук В.І., Молчанова М.О. // *Актуальні проблеми комп’ютерних наук : зб. наук. пр. за матеріалами XVII Всеукр. наук.-практ. конф. (АПКН-2025).* – Хмельницький, 14–15 листоп. 2025 р. – Хмельницький, 2025. – С. 15–19.

2. Молчанова М.О., Андрощук В.І., Шурипа М.О., Залуцька О.О., Мазурець О.В. Об’єктно-орієнтований підхід до нейромережевого виявлення суб’єктів кібербулінгу за повідомленнями у керованому хмарному середовищі / М.О. Молчанова, В.І. Андрощук, М.О. Шурипа, О.О. Залуцька, О.В. Мазурець // *Науковий журнал «Системи та технології».* – Суми, 2026. – № 1 (Прийнято до друку).

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

за матеріалами XVII Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2025»

14-15 листопада 2025

Хмельницький 2025

ЗМІСТ

УДК 004.8

Андрошук В.І., Молчанова М.О.

Хмельницький національний університет

ТРАНСФОРМЕРНЕ ВИЯВЛЕННЯ СУБ'ЄКТІВ КІБЕРБУЛІНГУ ЗА ТЕКСТОВИМИ ПОВІДОМЛЕННЯМИ

Розглянуто підхід до трансформерного виявлення кібербулінгу, орієнтований на інтерпретоване визначення не лише факту агресивної комунікації, а й суб'єктів впливу та спрямованості взаємодії. На першому рівні трансформерна модель класифікує повідомлення щодо наявності ознак кібербулінгу, після чого за пороговим значенням активується модуль залежнісного синтаксичного аналізу та семантико-рольової інтерпретації. Для навчання використано спеціалізовані корпуси кібербулінгу зі збалансованим розподілом класів і стилістичною різнорідністю даних. Реалізований прототип інтелектуальної системи забезпечує автоматизоване виявлення кібербулінгу та візуалізацію рольових зв'язків, що підвищує пояснюваність результатів і придатність рішення до інтеграції в системи цифрової безпеки.

The paper presents the transformer-based approach to cyberbullying detection focused on interpretable identification of both aggressive communication and its underlying actors and targets. At the first level, a transformer model classifies messages for the presence of cyberbullying indicators; once a probability threshold is exceeded, a dependency-based syntactic and semantic role analysis. The neural component is trained on specialized cyberbullying corpora combining balanced class distributions with stylistically diverse data. The implemented prototype of an intelligent system provides automated cyberbullying detection and visualization of role relations, improving the explainability of results and supporting integration into digital safety and social media monitoring systems.

Агресивна взаємодія в цифрових комунікаціях дедалі частіше набуває непрямих форм, у яких кібербулінг проявляється не стільки через відверто образливу лексику, скільки через серії натяків, приниження, остракізм і рольові «підсилювачі» конфлікту [1, 2]. Більшість наявних систем обмежуються бінарною класифікацією «токсично / не токсично» на рівні окремих повідомлень і не відтворюють суб'єктну структуру дискурсу: хто ініціює агресію, на кого вона спрямована, хто її підтримує або транслює далі [3]. Це знижує інтерпретованість результатів автоматизованого моніторингу та ускладнює практичне використання моделей у модераційних і превентивних системах [4].

Отож, зростання обсягів цифрової комунікації та перенесення значної частини соціальної взаємодії у мережеві середовища зумовлює необхідність автоматизованого моніторингу проявів кібербулінгу [5, 6]. У соціально-орієнтованих сервісах агресивні повідомлення поширюються з високою швидкістю,

формуючи середовище підвищеного ризику для вразливих груп, особливо підлітків [7]. Традиційні методи модерації виявляються недостатньо ефективними через масштабність потоків даних [8], багатомовність [9], контекстуальну мінливість [10] і постійну еволюцію мовних патернів [11], у яких агресія може маскуватися під сарказм, іронію чи непрямі форми впливу [12]. Тому сучасні дослідження у сфері Natural Language Processing орієнтовані на побудову більш гнучких і семантично чутливих моделей [13], здатних виявляти не лише факт вербальної агресії [14, 15], а й структуру комунікативної взаємодії [16].

У цьому контексті трансформерні моделі демонструють суттєві переваги завдяки механізму уваги, що дає змогу моделі фокусуватися на ключових словах і міжсловних залежностях, релевантних для інтерпретації агресивної поведінки [17]. На відміну від класичних підходів, що ґрунтувалися на частотних ознаках або поверхневій лінгвістичній структурі, трансформери здатні враховувати широкий контекст висловлювання й моделювати латентні семантичні зв'язки, характерні для складних соціально-комунікативних патернів [18]. Це забезпечує підвищену точність розпізнавання прихованих або непрямих форм кібербулінгу, а також сприяє кращій генералізації на даних з різних платформ [19].

Актуальною науковою задачею є також ідентифікація суб'єктів кібербулінгу – визначення того, хто є ініціатором агресивної дії і хто зазнає її впливу [20]. Таке завдання виходить за межі класичної бінарної класифікації й потребує глибшого аналізу синтаксичних та семантичних структур тексту. Сучасні NLP-підходи інтегрують трансформери з методами залежнісного аналізу, семантико-рольового розподілу та векторизації іменованих сутностей, що уможливлює побудову пояснюваних структур взаємодії. Виявлення ролей учасників є критичним для застосування в освітніх або соціальних системах, де реакція на інцидент має базуватися не лише на самому факті агресії, а й на коректному визначенні її джерела та адресата [21].

Суттєвий потенціал для розвитку мають і мультимодальні підходи, які поєднують текстовий аналіз з метаданими, реакціями користувачів, часовими патернами та специфікою платформи. Такі системи дозволяють враховувати ширший контекст комунікації, що часто є ключовим для точного розрізнення конфлікту, іронії та кібербулінгу [22]. Додатково, перспективним є застосування моделей із вбудованими пояснювальними механізмами, здатних генерувати прозорі аргументи щодо виявленої агресії, що є необхідною умовою інтеграції таких систем у регуляторні та безпекові платформи.

Загалом, розвиток NLP у сфері виявлення кібербулінгу визначається потребою у моделях, що поєднують високу точність, контекстну чутливість та пояснюваність. Трансформерні архітектури відкривають можливість створення комплексних систем, здатних аналізувати не лише зміст повідомлення, а й структуру комунікативної взаємодії між користувачами. Такі підходи формують потенційну основу для інтегрованих систем цифрової безпеки, орієнтованих на

раннє виявлення, превенцію та аналіз ризикової комунікації у масштабних онлайн-платформах.

Метою роботи є підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом переходу від виявлення самого факту агресивної комунікації до ідентифікації суб'єктів впливу та спрямованості взаємодії на основі трансформерних моделей. Об'єктом дослідження є процес автоматизованого виявлення кібербулінгу та його суб'єктів у текстових комунікаціях, предметом – моделі, методи та програмні засоби обробки природної мови для інтерпретованого аналізу агресивних висловлювань.

Запропонований метод реалізує багаторівневу обробку тексту. На першому рівні трансформерна модель виконує класифікацію повідомлень щодо наявності ознак кібербулінгу, формуючи ймовірнісну оцінку належності висловлювання до агресивного контенту. Після перевищення заданого порогу текст переходить до другого рівня аналізу, де поєднуються залежнісний синтаксичний розбір і семантико-рольова інтерпретація. Для кожного речення відновлюється предикативна структура, виділяються предикати та їх актанти, на основі чого текстове повідомлення відображається у множину семантичних трійок виду «суб'єкт – дія – об'єкт». У результаті для кожного фрагмента формується пара ⟨y, R⟩, де y – індикатор наявності кібербулінгу, а R – структура рольових зв'язків між учасниками комунікації.

Методологічно це дозволяє відокремити два логічні шари: детекцію агресії як такої та інтерпретацію комунікативних ролей у межах виявлених токсичних висловлювань. Використання dependency-parsing і механізмів відновлення актантних ролей забезпечує автоматизоване виокремлення потенційних кривдників і жертв, а також проміжних суб'єктів, які підсилюють або транслюють агресію, включно з випадками, коли вплив реалізується через конструкції з копулою чи непрямі характеристики, а не через явно образливі дієслова. Це особливо важливо для соціальних платформ, де значна частина кібербулінгу має завуальований характер.

Для навчання та валідації нейромережевої компоненти використано спеціалізовані корпуси кібербулінгу з відкритих джерел, що містять марковані приклади агресивних і нейтральних повідомлень, а також підмножини з деталізацією типів ворожих висловлювань. Один із застосованих корпусів характеризується збалансованим розподілом класів за основними категоріями кібербулінгу, інший – стилістичною й тематичною різнорідністю, оскільки агрегує дані з різних платформ. Комбіноване використання цих наборів даних дозволяє одночасно забезпечити стабільність базової класифікації та підвищити здатність моделі узагальнюватися на різні типи дискурсу.

На основі запропонованого методу спроєктовано та реалізовано прототип інтелектуальної системи, що поєднує трансформерний класифікатор, синтаксико-семантичний аналізатор та веб-інтерфейс для інтерактивного аналізу текстів. Система забезпечує введення довільних повідомлень, автоматичне визначення

наявності кібербулінгу, формування відсоткової оцінки ризику і візуалізацію виявлених суб'єктно-об'єктних зв'язків у вигляді таблиці семантичних трійок. Це створює передумови для інтеграції рішення в інформаційні системи цифрової безпеки, освітні платформи та аналітичні модулі моніторингу соціальних мереж.

Узагальнюючи, розроблений підхід демонструє можливість переходу від плоскої детекції токсичності до структурованого опису кібербулінгових ситуацій з урахуванням ролей учасників і спрямованості впливу. Багаторівневе поєднання трансформерних моделей і синтаксико-семантичного аналізу підвищує пояснюваність результатів і придатність системи до практичного використання. Подальші дослідження доцільно спрямувати на розширення рольової таксономії, адаптацію методу до україномовних і багатомовних корпусів та інтеграцію часово-графових моделей, що дозволить аналізувати динаміку кібербулінгу на рівні довготривалих комунікативних сценаріїв.

Перелік посилань

1. Civila S. Cyberbullying. Comprehensive Sexuality Education for Gender-Based Violence Prevention. 2024. P. 229–245.
2. Casas F. Age Discrimination. Encyclopedia of Quality of Life and Well-Being Research. Cham, 2023. P. 118–121.
3. Lee H. Lived Religion in Religious Vaccine Exemptions. Perspectives in Biology and Medicine. 2024. Vol. 67, no. 1. P. 96–113.
4. Протидія булінгу. МОН. URL: https://mon.gov.ua/tag/protidiya-bulingu?&type=all&tag=protidiya-bulingu
5. Антибулінг. AIKOM. URL: https://aikom.iea.gov.ua/bullying/help
6. Unnava S., Parasana S. R. A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach. Engineering, Technology & Applied Science Research. 2024. Vol. 14, no. 4. P. 15607–15613.
7. Mazurets O., Vit R. Practical Application of Method of Thematic Classification of Text Information Using LDA. Information Technology and Implementation (Satellite). Proceedings 11th International Conference. November 21, 2024. Kyiv, Ukraine. 2024. Pp. 151-152.
8. Віт Р.В., Мазурець О.В. Тематична класифікація текстової інформації засобами обробки природної мови. Збірник наукових праць XXIII Міжнародної наукової конференції «Нейромережні технології та їх застосування НМТі3-2024». 11-12 грудня 2024. Краматорськ-Тернопіль, ДДМА. 2024. с. 63-66.
9. Овчарук О.М., Мазурець О.В. Нейромережеве діагностування проявів ПТСР у текстовому контенті з використанням помилко-орієнтованого навчального набору даних. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №6, Т.1 (343). С. 195-200.
10. Крак Ю.В., Дідур В.О., Молчанова М.О., Мазурець О.В., Собко О.В., Залуцька О.О., Бармак О.В. Метод виявлення політичної пропаганди в інтернет-контенті нейромережевими засобами обробки природної мови. Науковий журнал «Проблеми програмування». Київ, 2024, №2-3. с. 288-295.
11. Овчарук О.М., Мазурець О.В. Нейромережева архітектура з квантовим шаром для аналізу текстових повідомлень на прояви посттравматичного стресового розладу. Науковий журнал «Наука і техніка сьогодні». Київ, 2024. №13 (41). С. 1192-1204.

12. Мазурець О.В., Тимофієв І.А., Кліменко В.І., Тищенко О.О. Метод виявлення депресивного стану пов'язаного із навчанням у закладах освіти із використанням нейромережі дуальної архітектури. Науковий журнал «Вісник Херсонського національного технічного університету». 2024. №4 (91). С. 311-318.
13. Віт Р.В., Мазурець О.В. Підхід до тематичної класифікації текстової інформації засобами обробки природної мови. Науковий журнал «Наукові праці Донецького національного технічного університету», серія «Проблеми моделювання та автоматизації проектування». 2025. №1 (21). С. 94-99.
14. Овчарук О.М., Мазурець О.В. Нейромережевий метод діагностування психологічних розладів за аналізом повідомлень на основі роздільного підходу до класифікації. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 1, 2025. с. 210-216.
15. Murava V., Zalutska O., Didur V., Mazurets O. Software architecture of information system for exchanging LLM thematic prompts. Global Trends in the Development of Information Technology and Science. Proceedings IV International Scientific and Practical Conference. June 25-27, 2025. Stockholm, Sweden. Pp. 121-127.
16. Юрченко Д.Ю., Овчарук О.М., Мазурець О.В., Шевчук П.О. Метод використання нейромережі гібридної архітектури для визначення емоційної тональності текстових повідомлень. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 2, 2025. с. 136-141.
17. Віт Р.В., Мазурець О.В. Метод виявлення психологічного цифрового перевантаження за аналізом текстових даних нейромережевими моделями глибокого навчання. Науковий журнал «Вісник Херсонського національного технічного університету». 2025. №2 (93). Т. 2. С. 107-114.
18. Віт Р.В., Мазурець О.В. Метод виявлення комунікаційних об'єктів як індикаторів цифрової втоми. Інтелектуальний метод виявлення цільових об'єктів предметної області для класифікації текстової інформації. Матеріали XIII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2025». 24-26.09.2025. Одеса. 2025. С.119-121.
19. Собко О.В. Метод нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості. Вісник Херсонського національного технічного університету. 2024. № 4 (91). С. 342–348.
20. Krak, I., Sobko, O., Mazurets, O., Tymofiiev, I., Molchanova, M., Barmak, O. Method for Detecting and Classifying Cyberbullying in Text Content Using Neural Networks. Lecture Notes in Networks and Systems. Springer, Cham, 2025, vol. 1473, pp. 486–498.
21. Krak I., Sobko O., Molchanova M., Tymofiiev I., Mazurets O., Barmak O. Method for neural network cyberbullying detection in text content with visual analytic. CEUR Workshop Proceedings, 2025, vol. 3917, pp. 298-309.
22. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.

УДК 004.8

Молчанова М.О., доктор філософії, старший викладач кафедри комп'ютерних наук
Хмельницький національний університет
ORCID: 0000-0001-9810-936X
Андрощук В.І., студент кафедри комп'ютерних наук
Хмельницький національний університет
Шурипа М. О., студент кафедри комп'ютерних наук
Хмельницький національний університет
ORCID: 0009-0003-7025-4647
Залуцька О.О., асистент кафедри комп'ютерних наук
Хмельницький національний університет
ORCID: 0000-0003-1242-3548
Мазурець О.В., к.т.н., доцент, доцент кафедри комп'ютерних наук
Хмельницький національний університет
ORCID: 0000-0002-8900-0650

Об'єктно-орієнтований підхід до нейромережевого виявлення суб'єктів кібербулінгу за повідомленнями у керованому хмарному середовищі

Мета роботи полягає у формуванні та обґрунтуванні об'єктно-орієнтованого підходу до нейромережевого виявлення суб'єктів кібербулінгу за повідомленнями із поєднанням первинної детекції та подальшої синтаксико-семантичної інтерпретації у керованому хмарному середовищі. Запропоновано узгоджену архітектуру, у якій нейромережевий модуль виконує фільтрацію повідомлень на рівні «кібербулінг / не кібербулінг», після чого результати проходять залежнісний аналіз із реконструкцією рольових зв'язків «суб'єкт-дія-об'єкт». Об'єктна модель (класи повідомлення, речення, токени,

залежності, предикати, рольові трійки, підсумок) забезпечує прозору трасованість рішень, а кероване хмарне середовище – відтворюваність запусків і масштабованість експериментів.

Експериментальна перевірка показала ефективність первинної детекції: модуль на основі BERT досяг метрики F_1 = 0.98 у двокласовій постановці («кібербулінг» / «не кібербулінг»), що забезпечує достатній рівень відсіву нерелевантних повідомлень перед рольовим аналізом. На експертно перевіреному піднаборі якість рольової ідентифікації підтверджено узгодженими показниками: для визначення суб'єкта отримано значення 0.88, 0.86, 0.87 за Precision, Recall, F_1 відповідно, для об'єкта – 0.85, 0.83, 0.84, для дієслівного центру 0.91, 0.89, 0.90; точне відновлення трійки дало показник F_1 = 0.76. Міжекспертна узгодженість становила значення коефіцієнта Коена 0.82 при 87% повної згоди, що вказує на надійність еталонних міток і коректність процедури оцінювання. Для вирішення спірних випадків застосовувався допоміжний третій суддя (LLM) з фіксованою інструкцією; фінальні мітки визначалися правилом більшості.

Отримані результати показують, що запропонований підхід не лише фіксує факт агресивної комунікації, а й надає структуровану інформацію про її адресність, будучи відтворюваним і аудитованим у практичних умовах. Це створює підґрунтя для інтеграції у модераційні системи та подальшого розширення на корпуси з детальнішою рольовою розміткою й багатомовною підтримкою.

Ключові слова: *кібербулінг; трансформерні моделі; синтаксико-семантичний аналіз; об'єктно-орієнтоване проєктування; кероване хмарне середовище.*

Molchanova M.O., Androshchuk V.I., Shurypa M.O., Zalutska O.O., Mazurets O.V. Object-oriented approach to neural network-based detection of cyberbullying subjects from messages in a managed cloud environment

This paper proposes and substantiates an object-oriented approach to neural detection of cyberbullying actors in user messages that combines primary classification with subsequent syntactic–semantic interpretation in a managed cloud environment. The architecture integrates a neural module that filters messages at the "cyberbullying vs. non-cyberbullying" level, followed by dependency-based analysis that reconstructs role relations ("subject-action-object"). An explicit domain model (message, sentence, token, dependency, predicate, role triple, and result classes) ensures transparent traceability from raw text to interpretable outputs, while the controlled cloud setup provides reproducible runs and scalable experimentation.

The experimental evaluation shows that the BERT-based detector achieved an F_1 score of 0.98 for the binary task, providing a high-quality filter prior to role interpretation. On an expert-verified subset, role identification was consistent: for the subject role, precision, recall and F_1 were 0.88, 0.86 and 0.87, respectively; for the object role, 0.85, 0.83 and 0.84; and for the verbal head, 0.91, 0.89 and 0.90. Exact match of the full role triple yielded an F1 of 0.76. Inter-annotator agreement reached Cohen's $\kappa = 0.82$ with 87% full agreement, indicating reliability of the reference labels and soundness of the evaluation protocol; disagreements were resolved by majority voting with a third, instruction-constrained LLM judge in an auxiliary capacity.

Beyond reporting accuracy at the detection stage, the study contributes a controllable and auditable processing pipeline that delivers structured information on the addressee and direction of the communicative act, rather than a binary verdict alone. The object-oriented design clarifies component responsibilities and invariants, simplifies testing and extension, and supports deployment in settings where reproducibility and documented decision paths are required. The results suggest practical applicability for moderation workflows and provide a foundation for future work on corpora with explicit role annotations, improved handling of implicit roles and coreference, and broader multilingual coverage.

Key words*: cyberbullying; transformer models; dependency parsing; object-oriented design; managed cloud.*

Постановка проблеми. У сучасних цифрових комунікаціях кібербулінг набуває контекстно залежних і часто непрямих форм [1]. Переважна частина автоматизованих рішень обмежується бінарною класифікацією повідомлень на «агресивні/неагресивні», що не забезпечує встановлення адресності впливу – зокрема, ідентифікації ініціатора, цільової особи та характеру мовленнєвого акту [2]. Відсутність рольової інтерпретації ускладнює побудову адресних інтервенцій, аудит рішень і аналітичний супровід модерації [3].

Актуальною є науково-практична задача розроблення багаторівневого підходу, який поєднує визначення наявності кібербулінгу [4] із подальшою реконструкцією суб'єктно-об'єктних зв'язків у висловлюваннях [5]. З інженерного погляду така задача потребує чіткої об'єктно-орієнтованої моделі предметної області, у якій повідомлення, речення, предикати та рольові трійки репрезентуються як окремі сутності із визначеними відповідальностями та інваріантами. Об'єктно-орієнтований підхід покликаний забезпечити модульність, розширюваність і тестованість багаторівневої обробки [6], від нейромережевої детекції ознак кібербулінгу до синтаксико-семантичного аналізу й інтерпретації ролей учасників.

Додаткові вимоги висуває обчислювальний контекст: результати мають бути відтворюваними та масштабованими у керованому хмарному середовищі, придатному для експериментів і прискореного інференсу [7]. Отже, постає потреба в методі та програмній реалізації, які поєднують нейромережеву детекцію з лінгвістично вмотивованою рольовою інтерпретацією, спираються на об'єктно-орієнтоване моделювання домену й забезпечують умови для подальшого оцінювання якості, продуктивності та практичної придатності.

Аналіз останніх досліджень і публікацій. Сучасні корпусні ініціативи з автоматизованого аналізу агресивної комунікації демонструють перехід від бінарної детекції до ієрархічних та пояснювальних постановок. У межах OffensEval на базі таксономії OLID показано, що трансформерні моделі ефективні не лише для виявлення образливості, а й для подальшої категоризації та ідентифікації мішені. У першій ітерації завдання (2019) найкращі результати

становили: F1 = 0.829 для підзадачі A (детекція образливості), F1 = 0.755 для підзадачі B (таргетована або нетаргетована образа) та F1 = 0.660 для підзадачі C (ідентифікація мішені IND/GRP/OTH). Підходи на основі BERT переважали серед лідерів [8]. Пояснювальна компонента була розвинена в SemEval 2021 Task 5 Toxic Spans, де метою стало виокремлення токсичних фрагментів на рівні символів. Найкраща команда досягла character F1 = 70.83%, що підтвердило практичну реалізованість спан-рівневого раціоналізування рішень трансформерами. Водночас зафіксовано помітну варіативність якості: окремі системи на основі BiLSTM CRF або ToxicBERT демонстрували F1 на рівні 62.23%, що вказує на складність спан-детекції та чутливість до архітектури і налаштувань [9].

Дослідження 2023 і 2024 років засвідчили конкурентоспроможність донавчання сучасних трансформерів у мовно і ресурсно обмежених сценаріях. Для бенгальської мови повідомлялося про F1 = 0.87 із використанням Bangla BERT або Multilingual BERT [10]. У домені мікроблогів окремі інженерні рішення на локальних твіттер-корпусах досягали F1 = 0.91 за специфічних експериментальних умов, що підтверджує важливість доменної адаптації [11]. Огляди узагальнюють стабільну перевагу sentence або cross encoder трансформерів над традиційними моделями і підкреслюють значення урахування сеансового контексту для підвищення надійності детекції [12].

Для пояснюваних рішень широке застосування отримав датасет HateXplain, який поєднує клас мови ворожнечі, ціль висловлювання та раціоналі. Це створює умови для спільного навчання детекції і інтерпретації (мішень і фрагмент) та підвищує довіру до модераційних систем [13, 14].

Підсумовуючи, наявні результати окреслюють зрілість трансформерних підходів у бінарній і багаторівневій класифікації, а також у раціоналізації на рівні спанів. Недостатньо опрацьованими залишаються аспекти переходу від класифікації і пояснення до відновлення повної рольової структури висловлювань та надійної ідентифікації суб’єктів впливу, а також питання

відтворюваності і масштабування таких рішень у керованому хмарному середовищі. Саме ці елементи визначають подальший вектор дослідження.

Мета дослідження. Сформувати та обґрунтувати об'єктно-орієнтований підхід до нейромережевого виявлення суб'єктів кібербулінгу за повідомленнями, що поєднує первинне визначення наявності кібербулінгу із подальшою синтаксико-семантичною інтерпретацією і відновленням рольових зв'язків «суб'єкт-дія-об'єкт». Передбачено розроблення об'єктної моделі предметної області, опис алгоритмічних етапів обробки, реалізацію прототипу у керованому хмарному середовищі та експериментальне оцінювання якості і продуктивності з акцентом на відтворюваність і масштабованість.

Викладення основного матеріалу дослідження. Запропонований об'єктно-орієнтований підхід до нейромережевого виявлення суб'єктів кібербулінгу за повідомленнями реалізує послідовну обробку текстових даних у керованому хмарному середовищі (рисунок 1). На вхід подаються неструктуровані повідомлення, які розглядаються як потенційні носії агресивної комунікації. Перший етап передбачає нейромережеве визначення наявності кібербулінгу: модель класифікує висловлювання за ознаками образливості, цькування чи вербального тиску, що дозволяє виокремити лише ті фрагменти, які потребують подальшої інтерпретації.

Рис. 1. Схема об'єктно-орієнтованого підходу до нейромережевого виявлення суб'єктів кібербулінгу

На наступному етапі здійснюється синтаксико-семантичний розбір повідомлень з реконструкцією граматичної структури та визначенням предикатів і лексичних носіїв дії. У межах об'єктно-орієнтованого подання

домену відповідні сутності повідомлення, речення, предикат і рольова трійка використовуються для впорядкування результатів аналізу та їх подальшої інтерпретації.

Завершальний етап спрямований на встановлення суб'єктно-об'єктних відносин у висловлюваннях: визначається, хто ініціює мовленнєвий акт, на кого він спрямований і яким чином реалізується вплив.

Подана UML-модель (рисунок 2) формалізує об'єктно-орієнтоване подання задачі нейромережевого виявлення суб'єктів кібербулінгу у повідомленнях та відображає фактичну логіку реалізації. Вхідне повідомлення репрезентується класом «Message» із базовими метаданими; воно композиційно розчленовується на речення, кожне з яких містить токени й мережу залежностей, що задає синтаксичну структуру висловлювання. На цій структурі визначається предикат як ядро дії з фіксацією типу (дієслівний, копулярний або номінальний) і полярності. Така стратифікація рівнів, від тексту через речення і залежності до предиката забезпечує узгоджений перехід від поверхневої форми до змістової інтерпретації.

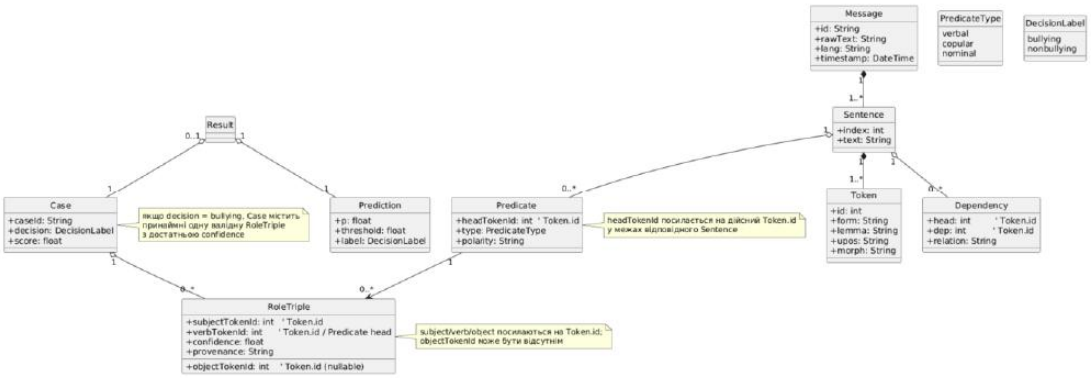


Рис. 2. UML-діаграма класів системи виявлення суб'єктів кібербулінгу за повідомленнями

Рольова реконструкція реалізується через клас «RoleTriple», який фіксує трійку «суб'єкт, дія, об'єкт» з посиланнями на відповідні токени, а також зберігає впевненість і походження отриманих зв'язків. За наявності кількох

актів мовленнєвого впливу трійки агрегуються у «Case», що містить підсумкове рішення та інтегральний бал і тим самим надає зручну одиницю для подальшого аналізу адресності. Первинне визначення факту кібербулінгу здійснюється нейромережевим класифікатором, результат якого акумулюється у класі «Prediction» у вигляді ймовірнісної оцінки, порогового значення та підсумкової мітки. Сукупний вихід конвеєра представлено класом «Result», який об'єднує рішення детектора з рольовою інтерпретацією у вигляді окремих трійок або цілісного випадку.

Запропонована об'єктна схема узгоджує дані і результати на всіх етапах обробки, дозволяє явно фіксувати посилання між рівнями представлення (від токенів до ролей) та забезпечує відтворюваність експериментів у керованому хмарному середовищі. Чітке розмежування сутностей і їхніх відповідальностей полегшує експериментальну валідацію, аудиту рішень і масштабування обробки, а також створює основу для подальшого оцінювання якості та продуктивності системи у прикладних сценаріях модерації.

Отримане структуроване подання випадку кібербулінгу поєднує факт агресії з чітко окресленою рольовою конфігурацією, що створює підґрунтя для аналізу адресності, динаміки взаємодій та подальшого практичного використання у керованому хмарному середовищі.

Запропонована архітектура виконання (рисунок 3) відображає розгортання підходу у керованому хмарному середовищі та структурована навколо єдиного Python-середовища, у межах якого функціонують основні компоненти обробки. Користувацький інтерфейс («UI/Demo») забезпечує подання вхідних повідомлень і перегляд результатів. Керування послідовністю етапів здійснює координатор обробки («Pipeline Coordinator»), який приймає повідомлення, ініціює первинне нейромережеве визначення наявності кібербулінгу, передає релевантні тексти на синтаксико-семантичний аналіз та збирає результати в узгодженому форматі.

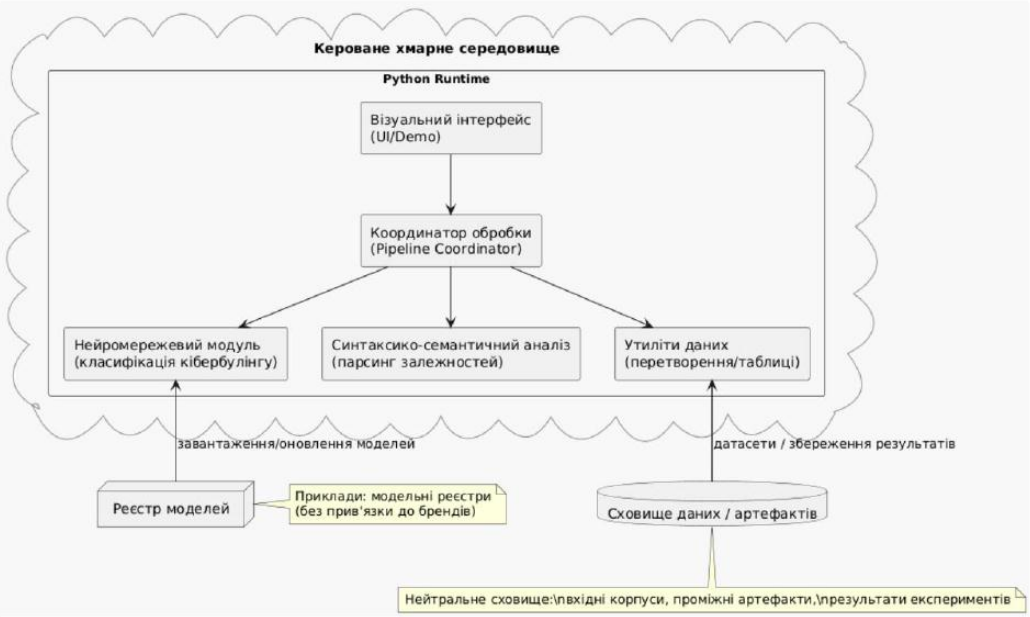


Рис. 3. Архітектура виконання у керованому хмарному середовищі

Нейромережевий модуль відповідає за класифікацію повідомлень щодо ознак кібербулінгу, повертаючи імовірнісне рішення, яке використовується як фільтр для подальшої інтерпретації. Блок синтаксико-семантичного аналізу виконує парсинг залежностей і формує підґрунтя для рольової реконструкції «суб’єкт-дія-об’єкт». Допоміжні утиліти даних здійснюють перетворення корпусів, формування підвибірок, кешування проміжних артефактів і підготовку підсумкових таблиць метрик, що спрощує відтворення експериментів.

Середовище взаємодіє з реєстром моделей, звідки здійснюється завантаження або оновлення модельних ваг та токенізаторів, і зі сховищем даних/артефактів, де розміщуються вхідні датасети, проміжні результати та підсумкові звіти. Така організація виключає прив’язку до конкретних сервісів, забезпечує контрольованість версій, можливість масштабування обчислень і прозоре трасування всіх кроків обробки від подання повідомлення до формування рішення про наявність кібербулінгу та отримання ролей учасників.

У роботі використано два відкриті корпуси: «Cyberbullying Classification» [15] та «Cyberbullying Detection» [16].

Корпус «Cyberbullying Classification» містить понад 47 тис. англомовних твітів із розміткою за шістьма категоріями: вікова, етнічна, гендерна, релігійна належність, інші прояви агресії та «не кібербулінг». Розподіл класів близький до збалансованого, що робить набір зручним для навчання і валідації бінарного детектора (об'єднання всіх проявів агресії в позитивний клас). Корпус відображає природну мову мікроблогів (сленг, орфографічні варіації, короткі висловлювання), однак не містить розмітки суб'єктів і об'єктів, тож не придатний для кількісної валідації рольових трійок.

Набір об'єднує тексти з різних платформ і використовується для двокласової постановки «кібербулінг / не кібербулінг». Різнорідність джерел підвищує узагальнюючу здатність моделі в задачі первинної детекції, проте наявний дисбаланс класів і відсутня деталізація типів агресії чи її адресності. У цій роботі корпус застосовано як допоміжний для перевірки узагальнюваності детектора; кількісна оцінка рольової інтерпретації на ньому не проводилась через брак відповідної розмітки.

Через відсутність у використаних корпусах «золотого стандарту» для рольових зв'язків (хто ініціатор, на кого спрямовано дію, яка дія), валідацію результатів рольової інтерпретації здійснювали два галузеві експерти та GPT-5 [17] як допоміжний суддя. Такий підхід забезпечує поєднання фахової оцінки з відтворюваною автоматизованою перевіркою.

Експерти незалежно анотували для кожного тестового прикладу трійку суб'єкт-дія-об'єкт і робили вердикт щодо коректності відновлення ролей системою. Узгодженість фіксували показниками міжекспертної згоди (відсоток повної згоди). У випадках розбіжностей застосовували схему більшості: два з трьох суддів формували остаточне рішення. GPT-5 використовували як третю сторону з чітким інструктажем, що вимагав: (1) вказати знайдені ролі, (2) навести опорні фрагменти тексту, (3) пояснити причину відхилення, якщо система помилилась. Вихід GPT-5 мав дорадчий характер і не замінював людське рішення.

У відкритих наборах даних, використаних для навчання детектора, відсутня рольова розмітка, тому потрібний людський еталон для перевірки відновлення суб'єктів. В свою чергу, залучення GPT-5 як додаткового судді підвищує відтворюваність і допомагає виявляти пропуски або неоднозначності, що залишилися поза увагою експертів. Також, така трирівнева схема скорочує витрати на повну ручну розмітку, зберігаючи при цьому якість і прозорість: фінальний вердикт завжди формується або повною згодою експертів, або більшістю з трьох суддів із фіксацією обґрунтувань.

У підсумку, обрана стратегія оцінювання дозволяє коректно перевіряти саме якість відновлення суб'єктів/об'єктів і дій, не обмежуючись лише метриками детекції кібербулінгу, та водночас відповідає вимогам відтворюваності, заявленим у меті дослідження.

Нейромережевий модуль первинної детекції реалізовано на основі трансформерного енкодера типу BERT [17] із донавчанням під двокласову постановку «кібербулінг» / «не кібербулінг». Передобробка обмежувалася стандартною токенізацією; усі прояви агресивної комунікації було зведено до позитивного класу. Оцінювання виконували на валідаційному піднаборі зі стратифікованим поділом даних. Запуски проводили у керованому хмарному середовищі з GPU класу T4, із фіксацією випадкових зерен і версій модельних ваг, що забезпечує відтворюваність без прив'язки до конкретної платформи.

Рольова інтерпретація ґрунтувалася на залежнісному аналізі речень і реконструкції трійок «суб'єкт-дія-об'єкт». За відсутності еталонної рольової розмітки кількісні метрики для цього етапу доповнювали контрактними перевірками узгодженості (валідність посилань на токени, єдиність предиката в реченні, несуперечливість ролей) та експертним рецензуванням, описаним у відповідному підрозділі. Така схема мінімізує залежність від конкретних реалізацій і водночас дає змогу коректно інтерпретувати отримані результати.

Нейромережевий модуль первинної детекції продемонстрував $F_1 = 0.98$ на валідаційних даних у двокласовій постановці «кібербулінг» / «не кібербулінг».

Через відсутність у відкритих корпусах «золотого стандарту» для ролей якість відновлення суб'єктів/об'єктів і предикатів оцінювали на експертно перевіреному піднаборі. У таблиці 1 наведено узагальнені показники (відносно узгоджених міток більшості).

Таблиця 1

Метрики для виявлення суб'єктів кібербулінгу

Завдання	Показник
Визначення суб'єкта	Precision / Recall / F_1: 0.88 / 0.86 / 0.87
Визначення об'єкта	Precision / Recall / F_1: 0.85 / 0.83 / 0.84
Визначення дієслівного центру	Precision / Recall / F_1: 0.91 / 0.89 / 0.90
Точне відновлення трійки	F_1: 0.76

Перед узгодженням рішень зафіксовано Cohen's $\kappa = 0.82$ при повній згоді 87%. Це свідчить про високу відтворюваність критеріїв анотування і достатню надійність отриманих оцінок. Використання GPT-5 як третього судді мало допоміжний характер: модель надавала пояснення й опорні фрагменти, що спрощувало вирішення спірних випадків і підвищувало прозорість процедури; фінальні мітки завжди визначалися правилом більшості.

Спеціалізований конвеєр забезпечує детерміноване й відтворюване відновлення рольових зв'язків «суб'єкт-дія-об'єкт» з фіксованими порогами та контрактними перевірками, тож результати є аудитованими і придатними для регламентного використання. Натомість вихід GPT-5 суттєво залежить від версії та формулювання підказки, що ускладнює формальну верифікацію та дотримання нормативних вимог.

Приклад роботи розробленого програмного забезпечення наведено на рисунку 4.

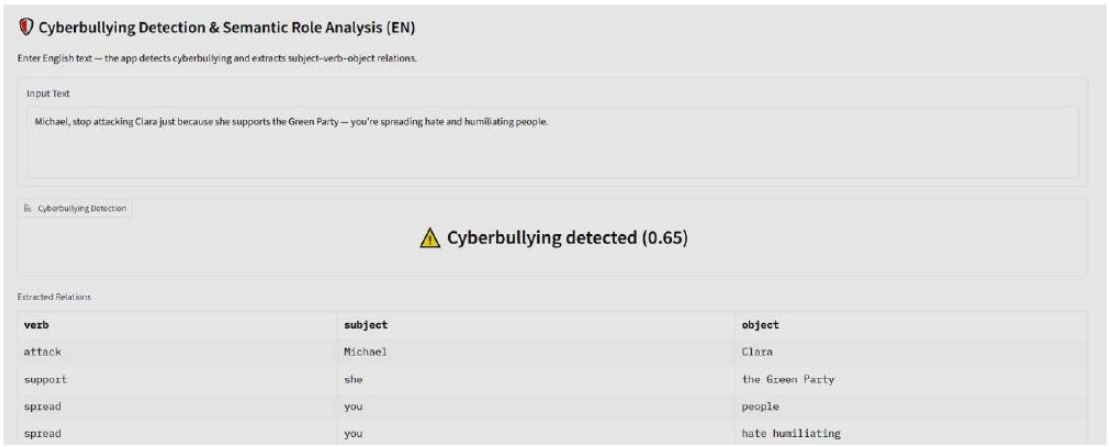


Рисунок 4 – Приклад роботи розробленого об’єктно-орієнтованого застосунку

Запропонований об’єктно-орієнтований підхід забезпечує узгоджену послідовність обробки: від нейромережевої детекції кібербулінгу до синтаксико-семантичної реконструкції зв’язків «суб’єкт-дія-об’єкт», що реалізовано у керованому хмарному середовищі з фіксацією параметрів для відтворюваності. Архітектура та процедура оцінювання створюють передумови для масштабованого застосування у модераційних сценаріях і подальшого розширення на корпуси з детальнішою анотацією ролей.

Висновок. В межах проведеного дослідження було досягнуто поставлену мету: сформовано та обґрунтовано об’єктно-орієнтований підхід, у якому нейромережеве виявлення кібербулінгу поєднано з подальшою синтаксико-семантичною інтерпретацією й відновленням рольових зв’язків «суб’єкт-дія-об’єкт», з реалізацією у керованому хмарному середовищі. У межах двокласової постановки первинний модуль на основі BERT продемонстрував F_1 на рівні 0.98, що підтверджує достатній рівень фільтрації повідомлень перед рольовим аналізом. На експертно-перевіреному піднаборі відтворення ролей для визначення суб’єкта зафіксовано значення 0.88, 0.86, 0.87 за метриками Precision, Recall, F_1 відповідно, для об’єкта 0.85, 0.83, 0.84, а для дієслівного центру 0.91, 0.89, 0.90, тоді як точне відновлення трійки дало F_1 у розмірі 0.76. Міжекспертна узгодженість становила к Коена 0.82 при 87 % повної згоди, що підтверджує надійність еталонних міток і коректність процедури оцінювання.

Сукупно результати свідчать, що запропонована архітектура забезпечує відтворюваність, аудитованість і практичну придатність: вона не лише фіксує факт агресивної комунікації, а й надає структуровану інформацію про адресність мовленнєвого впливу, створюючи підґрунтя для впровадження в модераційні системи та подальшого розширення на корпуси з детальною рольовою розміткою.

Список використаних джерел:

1. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms / S. Kim та ін. *Proceedings of the ACM on Human-Computer Interaction*. 2021. Т. 5, CSCW2. С. 1–34. URL: https://doi.org/10.1145/3476066 (дата звернення: 07.11.2025).

2. Paul S., Saha S., Hasanuzzaman M. Identification of cyberbullying: A deep learning based multimodal approach. *Multimedia Tools and Applications*. 2020. URL: https://doi.org/10.1007/s11042-020-09631-w (дата звернення: 07.11.2025).

3. Human Activity Recognition for the Identification of Bullying and Cyberbullying Using Smartphone Sensors / V. Gattulli et al. *Electronics*. 2023. Vol. 12, no. 2. P. 261. URL: https://doi.org/10.3390/electronics12020261 (дата звернення: 07.11.2025).

4. Method for neural network cyberbullying detection in text content with visual analytic / I. Krak et al. *CEUR Workshop Proceedings*. 2025. Vol. 3917, PP. 298-309. URL: https://ceur-ws.org/Vol-3917/paper57.pdf (дата звернення: 07.11.2025).

5. Method for cyberbullying neuronetwork detection using cloud services and object-oriented model / М. Молчанова та ін. *Herald of Khmelnytskyi National University. Technical sciences*. 2024. Vol. 333, no. 2. P. 200–206. URL: https://doi.org/10.31891/2307-5732-2024-333-2 (дата звернення: 07.11.2025).

6. Verma R., Kumar K., Verma H. K. Code smell prioritization in object-oriented software systems: A systematic literature review. *Journal of Software:*

Evolution and Process. 2023. URL: https://doi.org/10.1002/smr.2536 (дата звернення: 07.11.2025).

7. Load Balancing in cloud Environment: A State of-the-Art Review / Y. Lohumi et al. *IEEE Access*. 2023. P. 1. URL: https://doi.org/10.1109/access.2023.3337146 (дата звернення: 07.11.2025).

8. OffensEval 2023: Offensive language identification in the age of Large Language Models / M. Zampieri et al. *Natural Language Engineering*. 2023. Vol. 29, no. 6. P. 1416–1435. URL: https://doi.org/10.1017/s1351324923000517 (дата звернення: 07.11.2025).

9. SemEval-2021 Task 5: Toxic Spans Detection / J. Pavlopoulos et al. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Online. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.semeval-1.6 (дата звернення: 07.11.2025).

10. Sihab-Us-Sakib S., Rahman M. R., Forhad M. S. A., Aziz M. A. Cyberbullying detection of resource constrained language from social media using transformer-based approach. Natural Language Processing Journal. 2024. Vol. 9. Article No. 100104. URL: https://doi.org/10.1016/j.nlp.2024.100104 (дата звернення: 07.11.2025).

11. Aliyeva Ç. O., Yağanoğlu M. Deep learning approach to detect cyberbullying on twitter. *Multimedia Tools and Applications*. 2024. URL: https://doi.org/10.1007/s11042-024-19869-3 (дата звернення: 07.11.2025).

12. Yi P., Zubiaga A. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*. 2023. Vol. 36. P. 100250. URL: https://doi.org/10.1016/j.osnem.2023.100250 (дата звернення: 07.11.2025).

13. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection / B. Mathew et al. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021. Vol. 35, no. 17. P. 14867–14875. URL: https://doi.org/10.1609/aaai.v35i17.17745 (дата звернення: 07.11.2025).

14. Hate-speech-CNERG/hatexplain · Datasets at Hugging Face. *Hugging Face – The AI community building the*

future. URL: https://huggingface.co/datasets/Hate-speech-CNERG/hatexplain (дата звернення: 07.11.2025).

15. Cyberbullying Classification. *Kaggle*. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification (дата звернення: 07.11.2025).

16. Cyberbullying Detection. *Kaggle*. URL: https://www.kaggle.com/datasets/gbiamgaurav/cyberbullying-detection (дата звернення: 07.11.2025).

17. GPT-5. OpenAI. URL: https://openai.com/gpt-5/ (дата звернення: 07.11.2025).

18. BERT applications in natural language processing: a review / N. M. Gardazi et al. *Artificial Intelligence Review*. 2025. Vol. 58, no. 6. URL: https://doi.org/10.1007/s10462-025-11162-5 (дата звернення: 07.11.2025).

References:

1. Kim, S., et al. (2021). *A human-centered systematic literature review of cyberbullying detection algorithms. Proceedings of the ACM on Human-Computer Interaction,* 5(CSCW2), 1–34. https://doi.org/10.1145/3476066

2. Paul, S., Saha, S., & Hasanuzzaman, M. (2020). *Identification of cyberbullying: A deep learning-based multimodal approach. Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-020-09631-w

3. Gattulli, V., et al. (2023). *Human activity recognition for the identification of bullying and cyberbullying using smartphone sensors. Electronics,* 12(2), 261. https://doi.org/10.3390/electronics12020261

4. Krak I., et al. (2025). *Method for neural network cyberbullying detection in text content with visual analytic*. CEUR Workshop Proceedings, 2025, vol. 3917, 298-309. URL: https://ceur-ws.org/Vol-3917/paper57.pdf

5. Molchanova, M., et al. (2024). *Method for cyberbullying neuronetwork detection using cloud services and object-oriented model. Herald of Khmelnytskyi*

National University: Technical Sciences, 333(2), 200–206. https://doi.org/10.31891/2307-5732-2024-333-2

6. Verma, R., Kumar, K., & Verma, H. K. (2023). *Code smell prioritization in object-oriented software systems: A systematic literature review. Journal of Software: Evolution and Process.* https://doi.org/10.1002/smr.2536

7. Lohumi, Y., et al. (2023). *Load balancing in cloud environment: A state-of-the-art review. IEEE Access,* 1. https://doi.org/10.1109/access.2023.3337146

8. Zampieri, M., et al. (2023). *OffensEval 2023: Offensive language identification in the age of large language models. Natural Language Engineering,* 29(6), 1416–1435. https://doi.org/10.1017/s1351324923000517

9. Pavlopoulos, J., et al. (2021). *SemEval-2021 Task 5: Toxic spans detection.* In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.6

10. Sihab-Us-Sakib, S., Rahman, M. R., Forhad, M. S. A., & Aziz, M. A. (2024). *Cyberbullying detection of resource-constrained language from social media using transformer-based approach. Natural Language Processing Journal,* 9, 100104. https://doi.org/10.1016/j.nlp.2024.100104

11. Aliyeva, Ç. O., & Yağanoğlu, M. (2024). *Deep learning approach to detect cyberbullying on Twitter. Multimedia Tools and Applications.* https://doi.org/10.1007/s11042-024-19869-3

12. Yi, P., & Zubiaga, A. (2023). *Session-based cyberbullying detection in social media: A survey. Online Social Networks and Media,* 36, 100250. https://doi.org/10.1016/j.osnem.2023.100250

13. Mathew, B., et al. (2021). *HateXplain: A benchmark dataset for explainable hate speech detection.* In *Proceedings of the AAAI Conference on Artificial Intelligence,* 35(17), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745

14. Hate-speech-CNERG/hatexplain. (n.d.). *Datasets at Hugging Face – The AI community building the future.* Hugging Face.

https://huggingface.co/datasets/Hate-speech-CNERG/hatexplain (Accessed November 7, 2025)

15. Cyberbullying Classification. (n.d.). *Kaggle.* https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification (Accessed November 7, 2025)

16. Cyberbullying Detection. (n.d.). *Kaggle.* https://www.kaggle.com/datasets/gbiamgaurav/cyberbullying-detection (Accessed November 7, 2025)

17. OpenAI. (2025). *GPT-5.* https://openai.com/gpt-5/ (Accessed November 7, 2025)

18. Gardazi, N. M., et al. (2025). *BERT applications in natural language processing: A review. Artificial Intelligence Review,* 58(6). https://doi.org/10.1007/s10462-025-11162-5

Додаток К

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

МЕТОД БАГАТОРІВНЕВОГО ВИЯВЛЕННЯ СУБ'ЄКТІВ ВПЛИВУ КІБЕРБУЛІНГУ НА ОСНОВІ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ

Виконав:
студент групи КНм-24-1
Владислав АНДРОЩУК

Керівник:
Ph.D., ст. викл. кафедри КН
Марина МОЛЧАНОВА

Актуальність

Актуальність теми визначається зростанням частоти агресивної взаємодії у цифрових комунікаціях та недостатньою здатністю автоматизованих систем виявляти приховані форми впливу у текстовому середовищі. Кібербулінг все частіше проявляється не через відверто образливу лексику, а через непрямі мовленнєві дії, що формують тиск, підтримують ескалацію або задають тональність дискурсу. Більшість існуючих підходів зосереджені на класифікації повідомлень як токсичних або нейтральних, тоді як структура взаємодії між авторами цих повідомлень залишається поза увагою.

У таких умовах важливим стає аналіз того, як різні учасники текстової комунікації впливають на розвиток агресивного контенту незалежно від їхньої персональної ідентифікації. Застосування трансформерних моделей відкриває можливість урахування контексту, послідовності та функціональної ролі мовленнєвих актів, що дозволяє досліджувати вплив суб'єктів не на рівні особи, а на рівні текстової поведінки. Розроблення методу багаторівневого виявлення таких суб'єктів забезпечує перехід від ізольованого аналізу висловлювань до моделювання комунікативної динаміки, що підвищує ефективність моніторингу, інтерпретації та попередження кібербулінгових проявів.

Мета і задачі роботи

Об’єкт дослідження – процес автоматизованого виявлення кібербулінгу та його суб’єктів у текстових комунікаціях.

Предмет дослідження – моделі, методи та засоби обробки природної мови для виявлення кібербулінгу та його суб’єктів у текстових комунікаціях.

Метою кваліфікаційної роботи магістра є підвищення інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб’єктів впливу та спрямованості взаємодії на основі трансформерних моделей

Для досягнення поставленої мети слід вирішити такі **завдання**:

- дослідити сучасний стан області виявлення кібербулінгу;
- виконати огляд сучасних методів та засобів виявлення кібербулінгу та суб’єктів впливу;
- виконати аналіз наукових досліджень предметної області;
- розробити метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;
- виконати підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу;
- здійснити програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;
- виконати дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

Підхід до багаторівневого виявлення суб'єктів впливу кібербулінгу

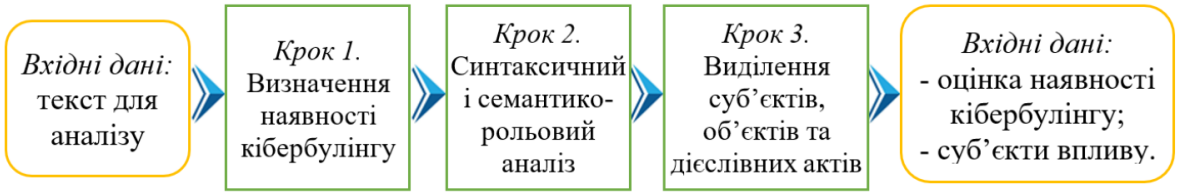
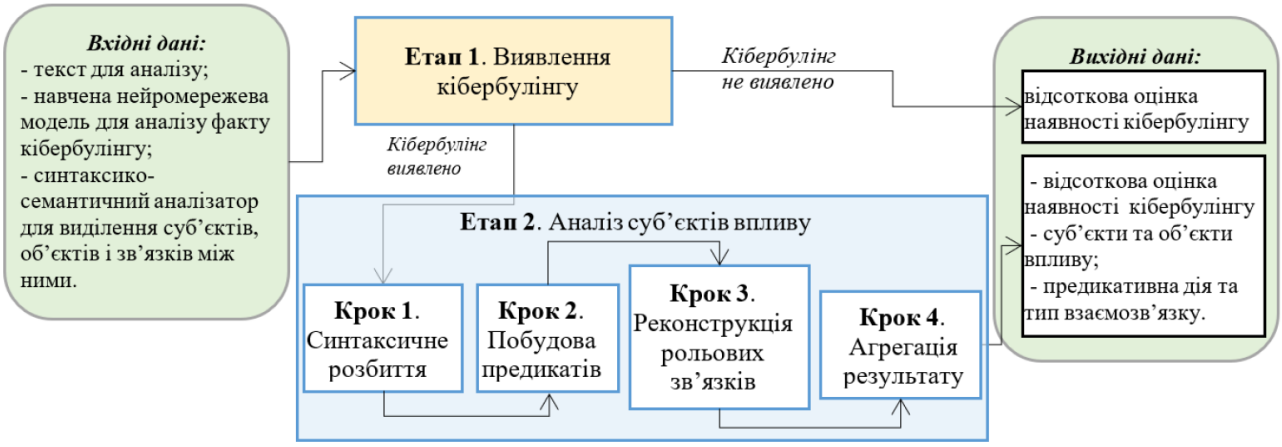
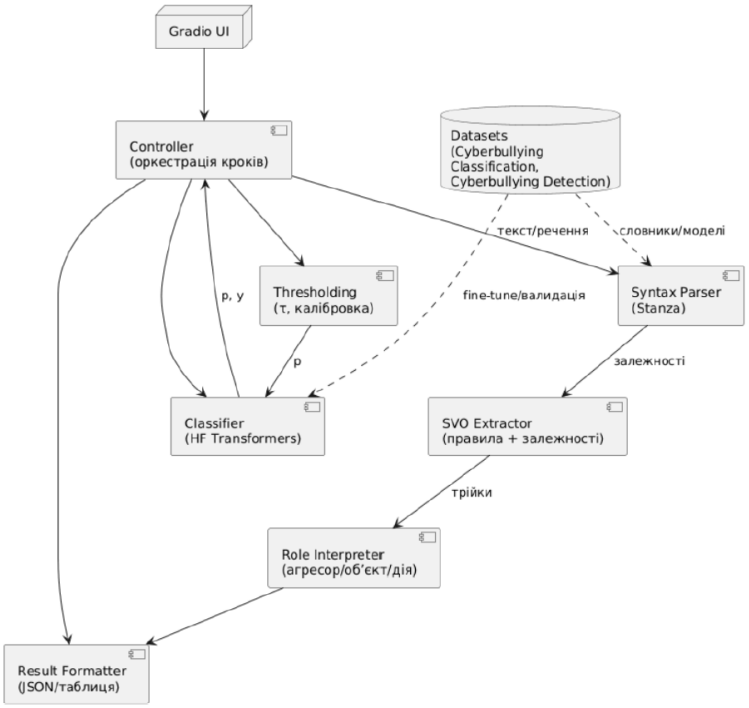


Схема методу багаторівневого виявлення суб'єктів впливу кібербулінгу



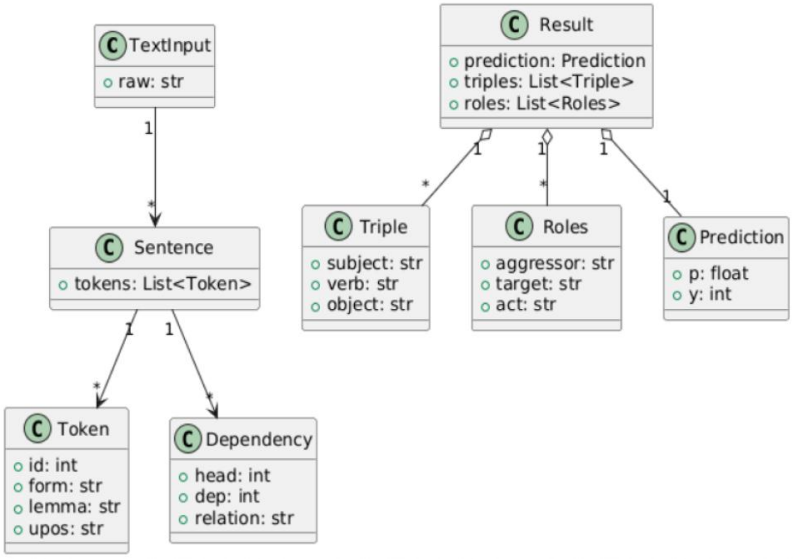
Датасет

1. Cyberbullying Classification
•Понад 47 тис. англомовних твітів
•6 категорій: за віком, етнічністю, гендером, релігією, іншими проявами агресії, «не кібербулінг»
•Збалансована вибірка (~8 тис. прикладів на категорію)
•Природна форма мовлення: сленг, короткі повідомлення, сарказм, орфографічні варіації
•Підходить для навчання моделей на виявлення кібербулінгу та його спрямування
•Обмеження: відсутня інформація про ролі суб'єктів і об'єктів впливу
2. Cyberbullying Detection
•Тексти з різних платформ: коментарі, обговорення, твіти
•Двокласова класифікація: кібербулінг / не кібербулінг
•Включає агресивні, токсичні, образливі або ворожі висловлювання
•Висока мовна та тематична варіативність → краща узагальнювальна здатність моделей
•Обмеження: дисбаланс класів, відсутність деталізації типу агресії та адресності
•Використання: базовий шар для загального виявлення токсичності або початкове донавчання нейромереж



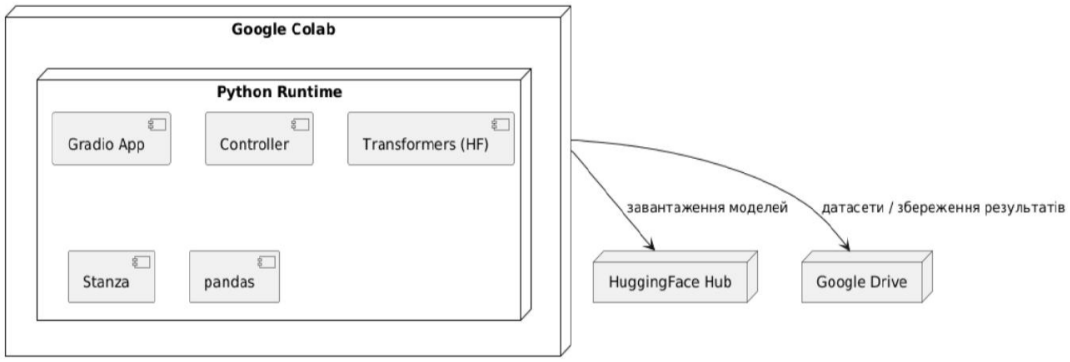
Проєктування складових інтелектуальної системи

Програмна структура компонентів інтелектуальної системи



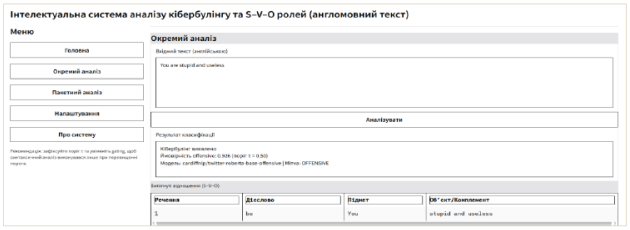
Діаграма класів інтелектуальної системи

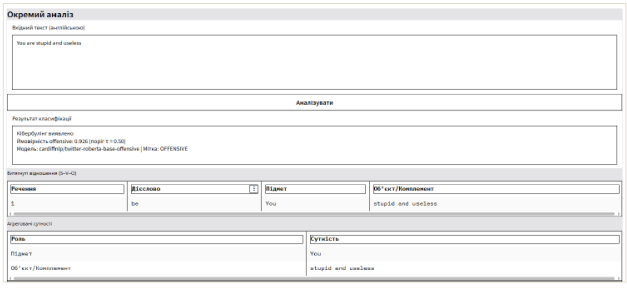
Програмна структура компонентів інтелектуальної системи

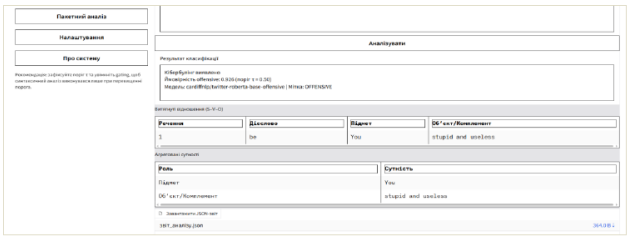


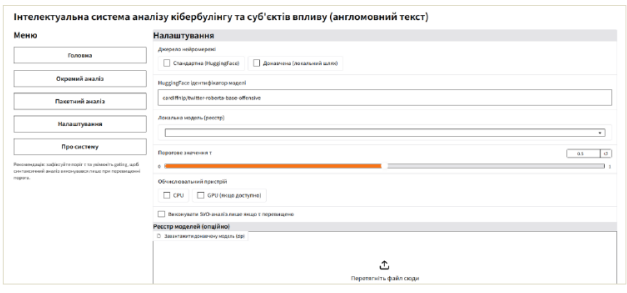
Архітектурна організація середовища

Інтелектуальна система









Дослідження методу

Динаміка навчання моделі DistilRoBERTa на зведеному корпусі

Епоха	Train loss	Val loss	Accuracy	Precision	Recall	F_1
2	0.312	0.338	0.921	0.914	0.907	0.910
3	0.279	0.324	0.929	0.923	0.916	0.920
4	0.256	0.318	0.934	0.928	0.922	0.925

Стійкість до доменного зсуву: якість на підвибірках різного походження (τ = 0.50), кількість зразків – 50.

Джерело повідомлень	Accuracy	Precision (кібербулінг)	Recall (кібербулінг)	F_1 (кібербулінг)
Twitter-повідомлення (короткий формат)	0.939	0.934	0.928	0.931
Інші платформи (коментарі, обговорення, повідомлення)	0.926	0.918	0.904	0.911

Дослідження методу

Результати на різних платформах та доменний зсув

•При переході до текстів з інших платформ спостерігається зниження метрик: Accuracy 0.926, F1 0.911

•Коментарі та обговорення мають довші висловлювання, складні дискурсивні зв'язки та контекстні відсилання → лексичні маркери агресії менш прямі

•Precision 0.918, Recall 0.904 → частина агресивних повідомлень залишається невиявленою, з'являються помилкові спрацювання

•F1 > 0.90 на вибірці з 50 зразків демонструє практичну придатність

•Для стабільної роботи в гетерогенних джерелах потрібні:

- доменна адаптація
- калібрування порога

Висновки

Було досягнуто мету кваліфікаційної роботи магістра, а саме інтерпретованості автоматизованого виявлення кібербулінгу шляхом виявлення не лише факту кібербулінгу, але й суб'єктів впливу та спрямованості взаємодії на основі трансформерних моделей.

Для досягнення поставленої мети було поставлено та вирішено такі завдання:

- досліджено сучасний стан області виявлення кібербулінгу;
- виконано огляд сучасних методів та засобів виявлення кібербулінгу та суб'єктів впливу;
- виконано аналіз наукових досліджень предметної області;
- розроблено метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;
- виконано підготовку навчальних даних для тонкої настройки нейромережі для виявлення кібербулінгу;
- здійснено програмну реалізацію методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей;
- виконано дослідження методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

ДЯКУЮ ЗА УВАГУ!

StrikePlagiarism.com

Документ прийнятий

Звіт подібності

Метадані

Назва організації	підрозділ
Khmelnytskyi National University	Кафедра комп'ютерних наук

Заголовок

КВАЛІФІКАЦІЙНА РОБОТА на тему Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей

Автор	Науковий керівник / Експерт
Владислав АНДРОЩУК	Марина МОЛЧАНОВА, Ph.D., ст. викл. кафедри КН

Кількість слів	Кількість символів	Дата звіту	Дата редагування	ІД документу
18317	154702	12/16/2025	12/16/2025	332880399

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

КП 1

18317
Кількість слів

КЦ

154702
Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		0
Інтервали		0
Мікропробіли		0
Білі знаки		1
Парафрази (SmartMarks)		37

Джерела

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз — Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://elar.khmnu.edu.ua/bitstreams/6279195e-75bb-4f3e-a900-7a18b3d791db/download	71 0.39 %
2	Стаття_СтТ [redacted] 11/11/2025 Publishing House "Helvetica" (Видавничий дім "Тельветика")	38 0.21 %

Tue Dec 16 09:49:40 EET 2025, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (UA) v-15.284 Educational

The maximum coincidence with one document 0.0%

Dictionaries check: en_US, ru_RU, ua_UA. **Errors in the documents: 14%**

	Document		Sum coincidence on the DB	
ID: 253199 Title: КВАЛІФІКАЦІЙНА РОБОТА на тему Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей Added in a DB: 2025-12-16 Authors: Владислав АНДРОЩУК Heads: Марина МОЛЧАНОВА Consultants: Opponents:	Symbols	Lexemes	Symbols	Lexemes
	139809	946	2276 (2%)	32 (3%)

Plagiarism sources

ID	Description	Plagiarism presence in the document	
		Symbols	Lexemes

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи *Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей*
Автор *студент групи КНм-24-1 Владислав АНДРОЩУК*
Освітня програма *Комп'ютерні науки*
Рівень вищої освіти *другий (магістерський)*
Спеціальність *122 – Комп'ютерні науки*
Науковий керівник: *Ph.D., ст. викл. каф. КН Марина МОЛЧАНОВА*

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмними засобами комісія зробила такий висновок:

№	Висновок	Позначка про відповідність
1	Ознаки академічного плагіату	
1.1	Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту.	***відповідає***
1.2	Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована.	
1.3	Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат.	
1.4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
2	Інші види порушень академічної доброчесності	***відсутні***

Підтвердження:

Запозичення, виявлені в роботі Владислава АНДРОЩУКА, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти, які не мають авторства і містять поширені конструкції та загальновідомі терміни, скорочення. Рівень подібності не перевищує допустимої межі. Таким чином, робота є законною та приймається до захисту.

Обсяг запозичень, визначений системами виявлення збігів/ідентичності/схожості:
- за системою Anti-Plagiarism: 2%;
- за системою StrikePlagiarism КП1: 4,2%, КП2: 1,1%.

16.12.2025

Завідувач кафедри _____ *Олександр БАРМАК*

Гарант освітньої програми _____ *Руслан БАГРІЙ*

Керівник кваліфікаційної роботи _____ *Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-24-1 Владислава АНДРОЩУКА за темою: *Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.*

1. Актуальність обраної теми

Проблема виявлення кібербулінгу та аналізу суб'єктів його впливу є надзвичайно актуальною у сучасних онлайн-середовищах. Автоматизовані методи багаторівневого виявлення агресивних повідомлень і визначення суб'єктів впливу дозволяють підвищити ефективність модерації, проводити соціально-аналітичні дослідження та забезпечувати безпечну комунікацію у соціальних медіа.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Кваліфікаційна робота повністю відповідає предметній області спеціальності 122 «Комп'ютерні науки». Вона демонструє застосування сучасних методів обробки природної мови, трансформерних моделей та нейромережевого аналізу текстових даних для вирішення прикладної задачі виявлення кібербулінгу.

3. Професійні та особистісні якості магістранта

Магістрант демонстрував аналітичне мислення та вміння систематизувати великі обсяги текстових даних. Під час виконання роботи проявлялася висока самодисципліна, відповідальність та здатність швидко опановувати сучасні методи трансформерного моделювання, що забезпечило успішне виконання всіх експериментальних та практичних завдань.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Всі ключові етапи роботи – від аналізу предметної області до розробки багаторівневого методу і побудови інтелектуальної системи – виконані самостійно. Магістрант проявив ініціативу у виборі архітектури моделей та налаштуванні параметрів нейромережі, що свідчить про високий рівень самостійності та професійної.

5. Наукова новизна та оригінальність запропонованих підходів

Наукова новизна роботи полягає у поєднанні трансформерних моделей із аналізом комунікативних ролей для багаторівневого виявлення кібербулінгу, що дозволило не лише

фіксувати факти агресії, а й визначати суб'єктів впливу та спрямованість їхніх взаємодій, підвищуючи пояснюваність і точність автоматизованого моніторингу.

6. Ступінь оволодіння методами дослідження

Магістрант продемонстрував здатність застосовувати складні алгоритми обробки природної мови, трансформерні моделі та методи багаторівневого аналізу комунікативних ролей. Він ефективно інтегрував різні методології для комплексного виявлення суб'єктів впливу кібербулінгу, що свідчить про високий рівень оволодіння сучасними науковими та прикладними методами дослідження.

7. Повнота та якість розкриття теми роботи

Тема роботи розкрита комплексно: досліджено відомі підходи до виявлення кібербулінгу, розроблено метод визначення суб'єктів впливу, побудовано інтелектуальну систему з інтерфейсом користувача та проведено її експериментальне тестування. Виконано аналіз етичних аспектів застосування, що підкреслює практичну та наукову цінність роботи.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Робота має логічну структуру та послідовний виклад матеріалу. Аргументація результатів базується на проведених експериментах і відповідає академічним стандартам. Літературна грамотність та оформлення рисунків і таблиць відповідають вимогам наукової роботи магістра.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Розроблений метод та інтелектуальна система можуть бути використані у практичних задачах модерації контенту, соціальної аналітики та моніторингу онлайн-комунікацій. Підхід дозволяє виявляти агресивні повідомлення, визначати суб'єктів впливу та спрямованість взаємодій, що робить його придатним для впровадження у соціальні медіа-платформи та дослідницькі проєкти.

10. Висновок про можливість допуску кваліфікаційної роботи до захисту, на яку оцінку заслуговує робота

Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «відмінно».

Науковий керівник _____ Ph.D., ст. викл. кафедри КН Марина МОЛЧАНОВА

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-24-1 Владислава АНДРОЩУКА за темою: *Метод багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.*

1. Актуальність обраної теми

Проблема кібербулінгу та його впливу на учасників цифрового середовища є надзвичайно актуальною в умовах зростання обсягів онлайн-комунікацій. Автоматизоване виявлення не лише фактів агресії, а й суб'єктів впливу дозволяє більш ефективно протидіяти кібербулінгу, проводити аналітику соціальних мереж і розробляти заходи для забезпечення безпечного цифрового середовища. Використання трансформерних моделей у цьому контексті підвищує точність та обґрунтованість отриманих результатів.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Кваліфікаційна робота магістра повністю відповідає предметній області спеціальності 122 «Комп'ютерні науки». У роботі застосовано сучасні методи обробки природної мови для вирішення прикладної задачі багаторівневого виявлення суб'єктів впливу кібербулінгу.

3. Повнота розкриття мети та завдань дослідження

Мета роботи та поставлені завдання виконані повністю. Автор дослідив сучасний стан проблеми, розробив метод багаторівневого виявлення суб'єктів впливу та розробив відповідну інтелектуальну систему, що дозволяє реалізувати практичні та експериментальні аспекти дослідження.

4. Наявність наукової новизни

Наукова новизна роботи полягає у поєднанні трансформерних моделей з аналізом комунікативних ролей для багаторівневого виявлення суб'єктів кібербулінгу. Такий підхід дозволяє визначати не лише факти агресії, а й суб'єктів впливу та напрямок взаємодії, підвищуючи пояснюваність результатів та відтворюючи структурну динаміку агресивних повідомлень.

5. Зміст кожного розділу роботи

Робота складається з чотирьох розділів. Перший розділ охоплює теоретичні засади комплексного виявлення кібербулінгу, аналіз наукових публікацій та постановку задачі. У другому розділі розроблено метод багаторівневого виявлення суб'єктів впливу на основі трансформерних моделей, описано датасет, критерії оцінки ефективності та етичні аспекти застосування. Третій розділ присвячено проєктуванню інтелектуальної системи, вибору інструментів, створенню інтерфейсу користувача, опису сценаріїв використання та розгортанню системи у хмарному середовищі. Четвертий розділ присвячений експериментальному дослідженні методу багаторівневого виявлення суб'єктів впливу кібербулінгу на основі трансформерних моделей.

6. Ступінь розкриття теми роботи

Тема кваліфікаційної роботи розкрита повністю. Проведено аналіз актуальності та існуючих методів, сформульовано завдання дослідження та розроблено прикладне рішення з перевіркою ефективності запропонованого підходу.

7. Якість оформлення кваліфікаційної роботи

Оформлення відповідає вимогам до наукових робіт магістра. Текст структурований, дотримано академічних норм викладу та оформлення рисунків і таблиць.

8. Недоліки кваліфікаційної роботи

До несуттєвих недоліків можна віднести обмежений обсяг експериментальної перевірки на різних корпусах даних та можливість додаткової оцінки стійкості моделі до більш складних доменних зсувів. Ці недоліки не впливають на загальну якість та наукову цінність роботи.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота

Кваліфікаційна робота магістра виконана на високому рівні та може бути допущена до захисту. Рекомендована оцінка – відмінно.

Опонент (прізвище, ім'я, по батькові, посада, місце роботи)

Бодрашов І.І. доц. каф. ІПЗ, ХНУ

«15» 12 2025 р

підпис