КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1189
УДК 004.056.5

Кольченко Віктор В'ячеславович
аспірант, асистент кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0002-0718-6859
viktor.v.kolchenko@lpnu.ua

Сабодашко Дмитро Володимирович
доктор філософії, старший викладач кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

Пятаєв Костянтин Костянтинович
студент кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0002-9646-7782
kostiantyn.piataiev.kb.2022@lpnu.ua

Городник Владислав Володимирович
студент кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0005-3169-0870
vladyslav.horodnyk.mkbas.2023@lpnu.ua

Щудло Іван Олексійович
аспірант кафедри інформаційно-вимірювальних технологій
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0003-0412-7682
ivan.o.shchudlo@lpnu.ua

Хома Юрій Володимирович
д.т.н., доцент кафедри інформаційно-вимірювальних технологій
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0000-0002-4677-5392
yurii.v.khoma@lpnu.ua

ДОСЛІДЖЕННЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ТОКСИЧНОГО КОНТЕНТ

Анотація. Стрімке зростання обсягів цифрової комунікації та поширення онлайн-платформ зумовили актуалізацію проблеми виявлення токсичного контенту, зокрема мови ненависті, кібербулінгу, погроз і дискримінаційних висловлювань. Такі прояви негативно впливають як на окремих користувачів, так і на інформаційне середовище загалом, підриваючи довіру до цифрових сервісів та сприяючи поширенню соціальних упереджень. У зв'язку з неможливістю масштабної ручної модерації особливої ваги набувають автоматизовані методи на основі глибинного навчання та обробки природної мови (NLP).

У роботі проведено порівняльний аналіз сучасних трансформерних моделей для виявлення токсичності тексту, зокрема Toxic BERT, Toxic Comment Model, RoBERTa Toxicity Classifier, DehaBERT Mono English, а також універсальної великої мовної моделі Phi-3 mini 4k без спеціалізованого донавчання. Оцінювання здійснено на двох відкритих наборах даних – Measuring Hate Speech та Jigsaw Toxic Comment Dataset – із використанням метрик Accuracy, Precision, Recall та F1-міри. Основний акцент зроблено на оптимізації F1-міри як збалансованого показника між прецизійністю та повнотою, а також застосовано механізм порогової фільтрації для коректного порівняння моделей із різною чутливістю до токсичності.

Результати експериментів засвідчили, що спеціалізовані трансформерні моделі, зокрема RoBERTa Toxicity Classifier, демонструють найвищу ефективність у виявленні явних форм токсичного мовлення, досягаючи точності понад 90% для категорій вираженої мови ненависті, погроз і закликів до насильства. Водночас встановлено зниження продуктивності для контекстно-залежних та непрямих проявів токсичності, таких як прихована агресія, неповага або образливий гумор. Показано, що дисбаланс класів у наборах даних істотно впливає на якість класифікації, зумовлюючи гірші результати для малопредставлених категорій. Також виявлено наявність упередженості моделей щодо окремих підгруп ідентичності та чутливість до використання ненормативної лексики.

Окремо продемонстровано, що застосування універсальної LLM без донавчання під конкретне завдання є малоефективним і ресурсозатратним під час інференсу. Отримані результати підтверджують доцільність використання спеціалізованих моделей для задач модерації контенту та вказують на перспективність комбінування універсальних великих мовних моделей із вузькоспеціалізованими класифікаторами з метою підвищення контекстно-залежного виявлення токсичності. Подальші дослідження доцільно спрямувати на розроблення ансамблевих підходів і зменшення модельної упередженості для забезпечення більш справедливих і надійних систем автоматизованої модерації.

Ключові слова: виявлення токсичності; мова ненависті; глибинне навчання; обробка природної мови (NLP); великі мовні моделі (LLMs).

ВСТУП

Зростаюча популярність онлайн-комунікації призвела до інтенсифікації випадків мови ненависті, кібербулінгу та переслідування, що створює ризики для окремих осіб і впливає на суспільство. Поширення шкідливого контенту порушує конструктивний діалог, підриває цифрову безпеку та закріплює негативні стереотипи. Окрім індивідуальної шкоди, толерування токсичності може зменшувати довіру до цифрових просторів, посилювати шкідливі упередження та сприяти радикалізації.

Контроль токсичності в соціальних мережах є одним із ключових викликів сучасного цифрового середовища. Ця проблема ускладнюється масштабами згенерованого контенту, динамічним характером взаємодій користувачів, а також лінгвістичною складністю й культурними нюансами комунікації. Додатковий рівень складності зумовлений використанням великих мовних моделей (LLMs), які можуть генерувати контент, що також містить упередженість і токсичність. Водночас сучасні технології штучного інтелекту відкривають нові можливості для автоматизованої модерації контенту, проте створюють і додаткові ризики, зокрема щодо упередженості та цензури [1], [2].

Наразі не існує загальноприйнятого визначення токсичності. Контент, який вважається токсичним, може істотно відрізнятися залежно від культури, віку, соціального контексту та суб'єктивного сприйняття реципієнта. Лінгвістичні конструкції, такі як іронія, сарказм або завуальовані образи, ускладнюють автоматизоване виявлення токсичної поведінки, що може призводити як до хибних спрацювань, так і до пропусків токсичності [2], [3].

Постановка проблеми. Щоденне зростання обсягів контенту робить ручну модерацію практично неможливою, що зумовлює необхідність використання автоматизованих систем виявлення токсичності. Методи на основі глибинного навчання, зокрема ті, що використовують LLMs, демонструють високу ефективність у таких завданнях, однак мають низку обмежень. Зокрема, ці моделі можуть відтворювати упередження, наявні в навчальних даних, що створює ризик як надмірної цензури, так і недостатнього контролю реально шкідливого контенту [4].

Таким чином, проблема контролю токсичності є багатовимірною та складною. Її ефективне розв'язання потребує цілісного підходу та співпраці між фахівцями з інформаційних технологій і гуманітарних наук, зокрема лінгвістики, соціології, етики, психології та права. Така міждисциплінарна взаємодія дає змогу застосовувати сучасні технології машинного навчання та штучного інтелекту з урахуванням соціального контексту, етичних принципів і прав людини, що зрештою сприяє створенню більш ефективних, точних і справедливих механізмів модерації контенту.

Аналіз літературних джерел показав, що основні підходи до виявлення та зменшення токсичності в онлайн-середовищі охоплюють такі напрями [5-7]:

- розроблення міжнародних законодавчих норм, які визначають відповідальність платформ за модерацію контенту та видалення токсичних матеріалів;
- підвищення цифрової грамотності користувачів, зокрема обізнаності щодо онлайн-токсичності, кібербулінгу та способів реагування на них;
- упровадження прозорих процедур модерації контенту й доступних механізмів оскарження рішень;
- використання інструментів на основі ШІ для аналізу контенту та підтримки модераторів у процесі ухвалення рішень;
- розроблення контекстно-залежних алгоритмів, здатних розпізнавати складні лінгвістичні нюанси та контекстуальні чинники з урахуванням мовних, культурних і соціальних особливостей;
- навчання моделей на багатомовних і репрезентативних наборах даних для зменшення упередженості та підвищення точності виявлення токсичності.

Інтеграція цих підходів сприятиме формуванню ефективніших механізмів протидії токсичності в соціальних мережах, забезпечуючи баланс між модерацією контенту та свободою вираження. Останні три пункти безпосередньо стосуються різних аспектів розроблення й використання автоматизованих інструментів моніторингу контенту та виявлення токсичності.

Аналіз останніх досліджень і публікацій. Перші автоматизовані інструменти виявлення токсичності ґрунтувалися на вручну сформованих правилах, лексиконах (списках слів) і застосуванні простих моделей машинного навчання, таких як Naïve Bayes або SVM як класифікаторів [8-10]. Такі системи є відносно простими в реалізації та ефективні для виявлення явно токсичного контенту, забезпечуючи базову функціональність модерації. Однак їхня результативність обмежена через високу ймовірність як хибних спрацювань, так і пропусків токсичності, що зумовлено нездатністю повноцінно враховувати контекст. Додатковим суттєвим недоліком цього підходу є потреба в постійному ручному оновленні лексиконів.

Вагомим проривом у розвитку автоматизованих систем виявлення шкідливого контенту стало впровадження технологій глибинного навчання. Для цього широко застосовувалися архітектури, такі як рекурентна нейронна мережа (RNN), мережа з довготривалою пам'яттю (LSTM), керований рекурентний блок (GRU) та згорткова нейронна мережа (CNN) [11], [12]. Завдяки здатності обробляти великі обсяги даних і навчатися складним шаблонам, моделі глибинного навчання аналізують текст на глибшому рівні, що дає змогу виявляти приховану токсичність. Водночас такі моделі мають і недоліки, зокрема потребу у великих розмічених наборах даних і високу обчислювальну складність.

У роботі [13] запропоновано гібридну архітектуру, що поєднує CNN і RNN для аналізу сентименту. Подібні гібридні підходи підвищують ефективність виявлення токсичності, поєднуючи здатність CNN ідентифікувати локальні патерни з можливістю

RNN захоплювати контекстуальні залежності в довгих послідовностях. Таку архітектуру також можна застосовувати для детекції токсичного контенту.

LSTM є спеціалізованою архітектурою RNN, розробленою для подолання проблеми згасання градієнта, що забезпечує ефективну обробку довгих послідовностей даних. Ключова перевага LSTM над класичними RNN полягає у здатності зберігати контекстну інформацію протягом тривалих часових інтервалів завдяки багаторівневій системі керування потоками інформації. У [14] LSTM використано як класифікатор для присвоєння міток «hate speech» або «non-hate speech» кожному вхідному твіту.

Модель GRU є ще одним різновидом RNN із довготривалою пам'яттю, проте, порівняно з LSTM, вона менш вимоглива до обчислювальних ресурсів. У дослідженні [15] порівняно продуктивність цих двох моделей за розміром навчального набору та метриками точності. За швидкістю навчання GRU приблизно на 30 % перевищує LSTM за обробки однакового набору даних, а також демонструє вищі показники точності, що робить її привабливою для задач із обмеженою кількістю навчальних прикладів, зокрема для національних мов.

Архітектура трансформерів здійснила революцію в NLP та суттєво покращила контроль токсичності на онлайн-платформах завдяки ефективнішій обробці складної лінгвістичної інформації та адаптації до різних сценаріїв використання. На відміну від RNN, які обробляють дані послідовно, трансформери забезпечують паралельну обробку всього вхідного тексту, що значно прискорює навчання. Крім того, вони краще моделюють контекстуальні залежності між словами завдяки механізму самоуваги та добре масштабуються на великих наборах даних і платформах [16].

Моделі на основі трансформерів, зокрема BERT (двобічний енкодер представлень із трансформерів) та його варіанти RoBERTa й DistilBERT, використовують специфічний підхід до навчання та обробки тексту для покращення його розуміння. У результаті вони здатні виконувати різні задачі класифікації тексту, включно з виявленням токсичності, і водночас залишаються відносно швидкими, що робить їх придатними для модерації контенту в реальному часі зі зменшенням упередженості [17].

Похідні від BERT моделі, такі як XLM-RoBERTa, забезпечують багатомовний аналіз, що особливо важливо для глобальних соціальних платформ. Такі моделі можуть додатково проходити донавчання для адаптації до нових форм токсичності або специфічних вимог окремих платформ [18].

Подальший розвиток автоматизованих систем контролю токсичності тісно пов'язаний із використанням LLMs, зокрема ChatGPT і Llama 2 [6, 19, 20]. Ці моделі також базуються на архітектурі трансформерів, однак, на відміну від BERT, орієнтованого переважно на розуміння тексту, LLMs здатні генерувати новий контент у відповідь на запити користувачів. У контексті виявлення токсичності застосовуються різні підходи: класифікація без додаткового навчання (Zero-Shot Classification), навчання з малою кількістю прикладів (Few-Shot Learning) та проєктування запитів (Prompt Engineering), які дають змогу адаптувати LLMs до задач модерації без повномасштабного донавчання [19, 21].

Метою статті є порівняльний аналіз сучасних трансформерних моделей для виявлення токсичності тексту та оцінювання їхньої ефективності на відкритих наборах даних із різними характеристиками.

Завданнями дослідження є:

1. Провести експериментальне порівняння спеціалізованих моделей виявлення токсичності та універсальної великої мовної моделі без донавчання;

2. Оцінити продуктивність моделей за метриками Accuracy, Precision, Recall та F1-score на різних наборах даних;

3. Проаналізувати вплив дисбалансу класів і контекстної складності на якість виявлення токсичного контенту.

Примітка. У цій статті наведено приклади образливого та токсичного мовлення. Фрази подано виключно з дослідницькою метою для ілюстрації роботи моделей з виявлення шкідливого контенту.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Огляд моделей. У сучасних умовах машинне навчання забезпечує найбільш ефективний підхід до виявлення токсичності в тексті. Тому для оцінювання сучасних методів детекції було використано найпоширеніші моделі цієї галузі. Усі розглянуті моделі глибинного навчання базуються на архітектурі трансформерів, що зумовлено її високою продуктивністю. Моделі обрано з урахуванням їхніх взаємодоповнювальних властивостей щодо виявлення як явних, так і прихованих форм токсичного контенту, що забезпечує стійкість до різнорідних текстових даних.

Використано чотири спеціалізовані моделі та одну універсальну LLM.

Toxic BERT. Класифікатор на основі BERT, донавчений на наборах даних токсичності. Підготовлені моделі та програмний код призначені для прогнозування токсичних коментарів у межах трьох змагань Jigsaw: Toxic Comment Classification, Unintended Bias in Toxic Comments та Multilingual Toxic Comment Classification [22]. У цьому дослідженні використано оригінальну модель (табл. 1) для коректнішого порівняння з іншими моделями, що формують бінарні результати.

Обмеження: якщо в коментарі присутні слова, асоційовані з лайкою, образами або ненормативною лексикою, імовірність його класифікації як токсичного є високою незалежно від тону чи наміру автора (наприклад, гумористичного або самоіронічного), що може спричиняти упередженість щодо вразливих груп населення [23].

Таблиця 1

Типи моделей TOXIC BERT [24]

Задача	Ціль	Першоджерело даних	Найкращий результат у рейтингу Kaggle	Оцінка детоксикації
Toxic Comment Classification Challenge	Створення багаторівневої моделі, що здатна виявляти різні типи токсичності, наприклад: загрози, нецензурну лексику, образи та ненависть на ґрунті ідентичності.	Wikipedia Comments	0.989	0.986
Jigsaw Unintended Bias in Toxicity Classification	Створення моделі, яка здатна розпізнавати токсичність і мінімізувати цей тип ненавмисної упередженості щодо згадок про ідентичність.	Civil Comments	0.947	0.936
Jigsaw Multilingual Toxic Comment Classification	Створення ефективних багатомовних моделей	Wikipedia Comments + Civil Comments	0.954	0.917

Toxic Comment Model. Ця модель є донавченою версією DistilBERT для класифікації токсичних коментарів. Вона орієнтована на модерацію онлайн-контенту, проте демонструє знижену якість для деяких підгруп ідентичності, зокрема згадок про мусульман. У табл. 2 наведено показники оцінювання для різних груп ідентичності.

Таблиця 2

Показник оцінювання для груп ідентичності [25] [26]

Підгрупа	Розмір підгрупи	AUC	BPSN AUC	BNSP AUC
Мусульмани	108	0.689	0.811	0.880
Євреї	40	0.749	0.860	0.825
Гомосексуали	56	0.795	0.706	0.972
Чорношкірі	84	0.866	0.758	0.975
Білі	112	0.876	0.784	0.970
Жінки	306	0.898	0.887	0.948
Християни	231	0.904	0.917	0.930
Чоловіки	225	0.922	0.862	0.967
Психологічні або психічні захворювання	26	0.924	0.907	0.950

AUC (Area Under the Curve) – площа під ROC-кривою, узагальнена метрика якості бінарної класифікації. BPSN AUC (Background Positive, Subgroup Negative AUC) характеризує здатність моделі відрізняти позитивні приклади фонової групи від негативних прикладів підгрупи ідентичності та використовується для виявлення хибних спрацювань щодо підгруп. BNSP AUC (Background Negative, Subgroup Positive AUC) оцінює розмежування негативних прикладів фону та позитивних прикладів підгрупи й відображає рівень пропусків токсичності.

RoBERTa Toxicity Classifier. Побудована на архітектурі RoBERTa [27] та оптимізована для нюансованого виявлення токсичності. Модель навчено на об’єднаному англомовному корпусі трьох наборів Jigsaw (2018, 2019, 2020), що містить близько 2 млн прикладів. Класифікатор демонструє AUC-ROC 0.98 та F1-міру 0.76 на тестовому наборі першого конкурсу Jigsaw [28].

DehateBERT-Mono-English. Спеціалізована версія BERT, донавчена для завдання виявлення мови ненависті з використанням наборів HateBase і Twitter Hate Speech Corpus. Такий фокусований процес навчання забезпечує підвищену точність ідентифікації hateful-контенту [29]. Модель є результатом роботи «Deep Learning Models for Multilingual Hate Speech Detection», прийнятої на ECML-PKDD 2020 [30].

Microsoft Phi-3 Mini. Phi-3-Mini-4K-Instruct – легка LLM-модель із 3,8 млрд параметрів, навчена на наборах Phi-3 (синтетичні дані та відфільтровані веб-ресурси з акцентом на якість і логічне міркування). Модель належить до сімейства Phi-3 і підтримує довжину контексту 4K токенів.

Післянавчальний етап включав контрольоване донавчання та оптимізацію прямих уподобань для інструктивного слідування й безпеки. За результатами бенчмарків (здоровий глузд, розуміння мови, математика, довгий контекст і логічне мислення) Phi-3 Mini-4K Instruct демонструє сучасний рівень продуктивності серед моделей до 13 млрд параметрів.

На відміну від попередніх моделей, Phi-3 Mini не навчалася спеціально для виявлення токсичності, а є універсальною мовною моделлю. Тому для експериментів застосовано спеціальний промпт:

```
    {"role":"system","content":"You are a model for identifying toxic content. Output must be 'toxic' or 'non toxic'"},
    {"role":"user","content":"Get the f out of here you j h son of a b h, I'll rape your f ing family if you don't leave you sem bastard. I will shoot you if you return, because you're a dirty sem , I hope you choke on a f ing bagel, prick. S om. We Came In?"},
    {"role":"assistant","content":"toxic"},
    {"role":"user","content":"Men are still ten times more likely to die on the job..."},
    {"role":"assistant","content":"non-toxic"}
```

МЕТОДИКА ДОСЛІДЖЕННЯ

Вибір джерел даних. Для оцінювання ефективності методів виявлення токсичності було обрано два відкриті набори даних на основі їх популярності, доступності документації та відповідних характеристик. Обидва містять тексти, створені людьми та розмічені як токсичні або нетоксичні.

Measuring Hate Speech Dataset. Набір містить понад 39 000 анотованих коментарів із Reddit, Twitter та YouTube. Основною змінною є «hate speech score», додатково доступні 10 ординальних міток (sentiment, disrespect, insult, humiliation, inferior status, violence, dehumanization, genocide, attack/defence, hate speech benchmark). Представлено 8 основних груп ідентичності та 42 підгрупи. Оцінка hate speech score є неперервною: >0.5 – приблизно мова ненависті, <-1 – контрмовлення або підтримка, -1…+0.5 – нейтрально або неоднозначно [31]. Деталі подано в табл. 3.

Таблиця 3

Теоретизовані якісні рівні спектра "мова ненависті – контрмовлення"

Рівень	Показник	Опис
5	Геноцид (Genocide)	Підтримка або намір систематично вбивати всіх або значну частину членів захищеної ідентичної групи
4	Насильство (Violence)	Загроза або застосування фізичної сили чи емоційного насильства з метою заподіяння шкоди або вбивства членів групи, що підлягає захисту
3	Дегуманізація (Dehumanization)	Позбавлення захищеної групи людських якостей, наприклад, порівняння з твариною, комахою або хворобою.
2	Ворожість (Hostility)	Недружелюбність або протидія захищеній групі, наприклад, через образи, нецензурну лексику або образи
1	Упередженість (Bias)	Схильність або перевага проти захищеної групи ідентичності, включаючи упередження
0	Нейтралітет (Neutral)	Описові або інші нешкідливі посилання на ідентичні групи
-1	Підтримка (Supportive)	Поважні, горді або інші повідомлення, що базуються на солідарності, про захищену групу (групи) ідентичності
-2	Контрмовлення (Counterspeech)	Реакція на мову ненависті, що має на меті підірвати її вплив та авторитет

У даному наборі даних представлено різні обсяги прикладів для окремих типів мови ненависті (рисунок 1). У супровідному дослідженні [32] було визначено прогнозовану складність кожної категорії (табл. 4), що загалом підтверджується подальшими науковими роботами.

Цей набір обрано як основний завдяки збалансованому представленню токсичних і нетоксичних прикладів.

Таблиця 4

Оцінка складності типів токсичності [32]

Тип	Складність
Настрій (Sentiment)	-2.62
Повага (Respect)	-2.26
Напад-Захист (Attack-Defend)	-1.10
Образа (Insult)	-0.94
Статус (Status)	-0.51
Дегуманізація (Dehumanize)	0.61
Приниження (Humiliate)	0.63
Мова ворожнечі(бінарна) (Hate speech (binary))	0.86
Насильство (Violence)	2.22
Геноцид (Genocide)	3.11

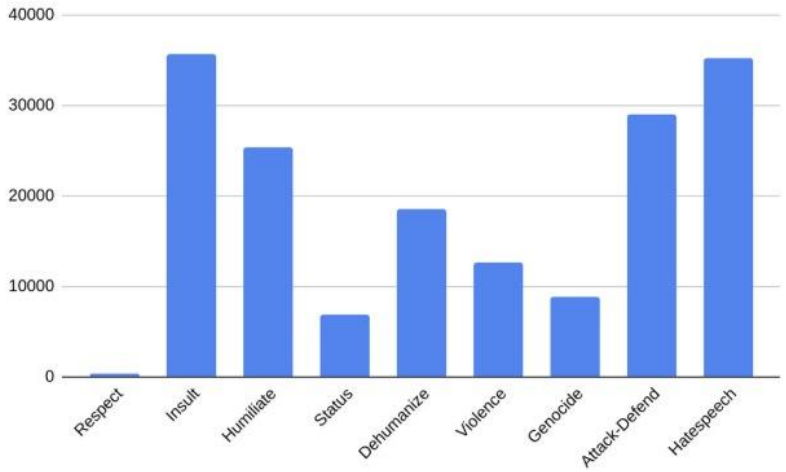

Рис. 1. Візуалізація кількості текстів із різними мітками в наборі даних Measuring Hate Speech

Jigsaw Toxic Comment Dataset. Широко використовуваний набір даних, що містить понад 200 000 розмічених коментарів, який забезпечує виявлення явної токсичності та атак за ознаками ідентичності. У межах цього змагання було сформовано два набори даних: один для навчання та один для валідації. У навчальному наборі стовпці, що характеризують текст, могли набувати двох значень: 0 і 1, що відповідали «ні» та «так». Тестовий набір подано окремо від його міток, і після їх об’єднання можливими стають три значення: -1, 0, 1, де -1 означає, що відповідний текст не використовується для валідації [33].

Оскільки багато моделей уже використовували цей набір для навчання, для оцінювання застосовано лише дані тестової вибірки. Після вилучення записів із міткою -1 залишалося 63 978 прикладів.

Основною проблемою цього набору даних є його дисбаланс. Із 63 978 записів 57 888 є нетоксичними, тоді як лише 6 090 − токсичними. Відповідно, це ускладнює формування коректних метрик і об’єктивне оцінювання моделей. На рис. 2 подано візуалізацію кількості токсичних текстів із різними мітками в цьому наборі даних.

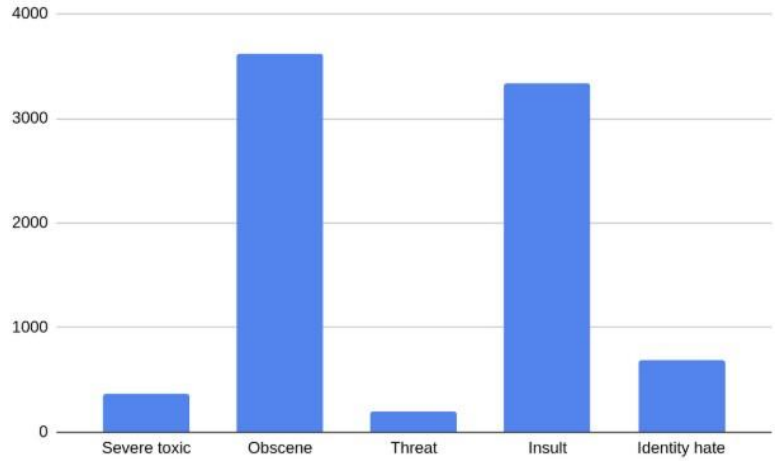

Рис. 2. Візуалізація кількості текстів із різними мітками в наборі даних Jigsaw Toxic Comment Dataset

Однак цей набір даних характеризується значним дисбалансом класів, де більшість прикладів є нетоксичними. З метою зменшення впливу цього чинника його було використано переважно для тестування та донавчання моделі, а не як основне джерело навчальних даних. Для реалізації цих цілей набір даних було поділено на окремі навчальну та тестову підвибірки.

Методологія. Ефективність системи виявлення токсичного контенту є ключовим параметром, що визначає її здатність коректно класифікувати вхідні дані. Для аналізу продуктивності таких систем застосовуються різні метрики, які забезпечують комплексне оцінювання з урахуванням як точності прогнозування, так і ресурсів, необхідних для його виконання. У межах цього дослідження використано такі показники: Accuracy (точність), Precision (прецизійність), Recall (повнота) та A1 Score (F1-міра).

У даному експерименті основною метою було оптимізувати F1-міру для кожної моделі як узагальнену метрику, що балансує між прецизійністю та повнотою. Такий підхід забезпечує досягнення компромісу між коректним виявленням токсичного контенту (Precision) та охопленням максимальної кількості реальних токсичних випадків (Recall).

Для подальшого підвищення якості моделей було впроваджено механізм порогової фільтрації (threshold-based filtering). Цей механізм дає змогу компенсувати надмірну або недостатню агресивність моделей під час класифікації токсичності. Наприклад, якщо модель прогнозує токсичність тексту зі значенням 0.85, але фактичний показник токсичності є нижчим за 0.85, відповідну мітку коригують на «non-toxic». Такий підхід дозволяє здійснювати об’єктивне порівняння моделей навіть у випадках, коли вони демонструють надто жорстку або, навпаки, надто м’яку поведінку щодо поточного набору даних.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати оцінювання моделей. Оцінювання запропонованої системи було зосереджено на аналізі продуктивності окремих моделей та ансамблевого підходу. Основним бенчмарком для тестування слугував набір даних Measuring Hate Speech завдяки його збалансованому представленню токсичних і нетоксичних прикладів.

Набір Jigsaw Toxic Comment Dataset (далі − Jigsaw Challenge), хоча й використовувався переважно для тестування, надав додаткові відомості щодо оброблення дисбалансу класів під час оцінювання моделей.

У результаті було сформовано 10 файлів результатів − відфільтровані набори даних разом із прогнозами моделей. На основі цих даних за допомогою методів бібліотеки scikit-learn обчислено метрики точності моделей, які наведено в табл. 5. Для кожної моделі порогове значення класифікації тексту як токсичного або нетоксичного було попередньо налаштовано з метою максимізації F1-міри, оскільки моделі можуть адаптуватися до специфічних вимог, а цей показник є збалансованим.

Таблиця 5

Результати оцінювання моделей

Jigsaw Challenge					
	Toxic BERT	Toxic comment model	RoBERTa Toxicity Classifier	DehaBERT Mono English	Phi-3 mini 4k
Precision	0.6991	0.5707	0.7207	0.4449	0.0627
Recall	0.7553	0.6913	0.8172	0.5271	0.3862
F1-score	0.7261	0.6252	0.7659	0.4825	0.1079
Accuracy	0.9458	0.9211	0.9525	0.8924	0.3920
Threshold setting	0.84	Positive > 0.64	Positive > 0.8	Positive > 0.15	–
Measuring Hate Speech					
Precision	0.8470	0.7458	0.8263	0.8255	0.6000
Recall	0.8689	0.6599	0.9239	0.7870	0.4357
F1-score	0.8578	0.7002	0.8724	0.8058	0.5048
Accuracy	0.8624	0.7302	0.8709	0.8188	0.5918
Threshold setting	0.86	Positive > 0.50	Positive > 0.98	Positive > 0.50	–

Як видно з таблиці, на обох наборах даних майже всі показники є найвищими для RoBERTa Toxicity Classifier; винятком є прецизійність для Measuring Hate Speech, де кращий результат продемонструвала Toxic BERT. Toxic BERT забезпечує стабільно високі результати на обох наборах, тоді як Toxic Comment Model і DehaBERT Mono English демонструють кращу продуктивність на одному наборі та гіршу − на іншому, що свідчить про їхню спеціалізацію на окремих типах контенту та обмеженість навчальних даних.

Візуальне порівняння основної метрики (F1-міри) наведено на рисунку 3.

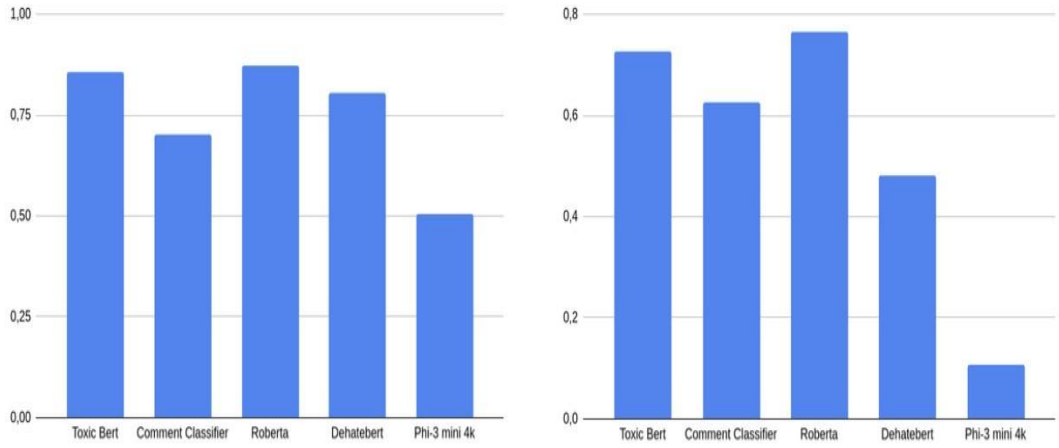

Рис. 3. F1-міра для Jigsaw Challenge (ліворуч) та Measuring Hate Speech (праворуч)

У табл. 6 наведено приклади того, як кожна з моделей оцінює токсичність текстів із набору даних Measuring Hate Speech.

Таблиця 6

Приклад виявлення токсичності

Категорія	Текст	Phi-3 mini 4k	Toxic BERT	Toxic comment model	RoBERTa Toxicity Classifier	DehaBERT Mono English	Hate Speech Score
Насильство (Violence)	Some of you bitches deserve to be used as human shields.	non-toxic	0.99	toxic / 0.95	toxic / 0.99	toxic / 0.86	4.01
Ворожість (Hostility)	Niggers are ruining this country	non-toxic	0.99	toxic / 0.93	toxic / 0.99	toxic / 0.69	2.32
Упередженість (Bias)	Okay taco munching beaner	non-toxic	0.32	non-toxic / 0.99	non-toxic / 0.99	toxic / 0.88	1.03
Підтримка (Supportive)	He doesn't know what sexuality means.	toxic	0.14	non-toxic / 0.99	non-toxic / 0.99	non-toxic / 0.97	-1.69
Контрмовлення (Counterspeech)	This picture is homophobic	toxic	0.72	non-toxic / 0.57	non-toxic / 0.99	non-toxic / 0.96	-2.36

Споживання ресурсів. Використання ресурсів і продуктивність є критично важливими чинниками під час вибору моделі для комерційного застосування, тому було проведено вимірювання основних параметрів. Обчислення виконувалися на ноутбуці з процесором Intel® Core™ i5-11320H 11-го покоління (3.20 GHz), графічним прискорювачем NVIDIA GeForce RTX 3050 Ti Laptop GPU та 16 ГБ оперативної пам’яті.

Проведення вимірювань для Microsoft Phi-3 mini 4k виявилося неможливим через її низькі результати під час оцінювання точності, а також неможливість запуску моделі на даному ноутбуці внаслідок нестачі необхідних обчислювальних ресурсів. Отримані результати наведено в табл. 7.

Таблиця 7

Вимірювання використання ресурсів

	Toxic BERT	Toxic comment model	RoBERTa Toxicity Classifier	DehaBERT Mono English
Середній час	2,40 ms	4,26 ms	3,27 ms	3,59 ms
Стандартне відхилення	4,91 ms	5,01 ms	4,96 ms	2,94 ms
Використання ОЗП	1,375,076	1,208,740	1,209,356	972,760

Ефективність виявлення токсичності для різних типів токсичного контенту. Для аналізу того, які аспекти мови ненависті модель розпізнає найкраще, було обрано RoBERTa Toxicity Classifier як очевидного лідера за показниками ефективності. З метою визначення продуктивності в окремих категоріях було обчислено загальну кількість токсичних коментарів у кожній категорії та кількість тих із них, які було коректно виявлено моделлю. На рисунках 4 і 5 наведено відповідні результати для обох наборів даних.

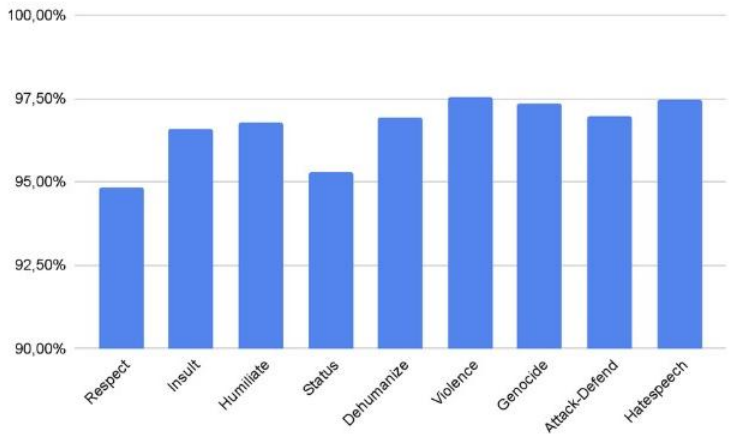

Рис. 4. Точність виявлення категорій токсичності в наборі даних Measuring Hate Speech

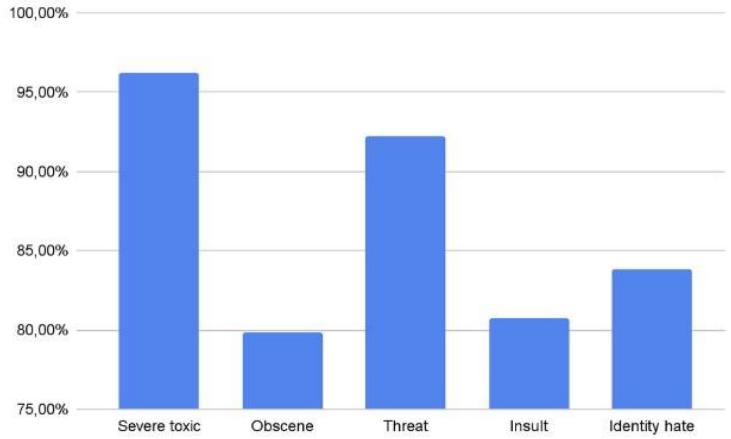

Рис. 5. Точність виявлення категорій токсичності в Jigsaw Challenge

У наборі даних Measuring Hate Speech для аналізу було використано токсичні коментарі з такими мітками, за якими рецензенти погоджувалися щодо наявності негативних характеристик (наприклад, insult, humiliate зі значеннями, що перевищують 3) або не погоджувалися з позитивними характеристиками (наприклад, respect, status зі значеннями, нижчими за 3). У Jigsaw Challenge застосовувалися виключно бінарні значення, надані набором даних.

На обох рисунках спостерігаються високі результати щодо виявлення очевидної мови ненависті: вираженої токсичності, погроз, hate speech, а також закликів до насильства або геноциду. Водночас відносно низькі показники зафіксовано для виявлення менш очевидних проявів агресії або ненормативного мовлення, зокрема obscene, insult, respect та status.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Проведене дослідження підтверджує, що спеціалізовані моделі на основі архітектури трансформерів, зокрема RoBERTa, є ефективними для виявлення явного токсичного контенту в наборах даних із різними характеристиками. Стабільна продуктивність Toxic BERT засвідчує надійність донавчених трансформерних архітектур у задачах автоматизованої модерації.

Порівняльний аналіз показав, що, хоча ці моделі добре розпізнають очевидну мову ненависті та погрози, їхня ефективність знижується для контекстно-залежних форм токсичності, таких як неявна упередженість або образливий гумор.

Використання універсальної мовної моделі без донавчання продемонструвало суттєві обмеження такого підходу, що свідчить про неможливість заміни спеціалізованих моделей без цільової адаптації. Збереження упередженості, особливо щодо контенту, пов’язаного з підгрупами ідентичності, підкреслює необхідність ретельного калібрування для уникнення несправедливих результатів модерації.

Подальші дослідження планується спрямувати на поєднання універсальних LLM зі спеціалізованими моделями з метою покращення контекстно-залежного виявлення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Guo, K., et al. (2023). An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)* (pp. 1568–1573). IEEE.
2. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *arXiv Preprint*. https://arxiv.org/abs/2009.11462
3. Biswas, P., & Haritha, D. (2024). Automatic hate speech detection and the hassle of offensive language. *Educational Administration: Theory and Practice, 30*(5), 12663–12668. https://kuey.net/index.php/kuey/article/view/4005
4. Hee, M. S., et al. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 4407–4419).
5. Prem, E., & Krenn, B. (2024). *On algorithmic content moderation*.
6. Roy, S., Harshavardhan, A., Mukherjee, A., & Saha, P. (2023). Probing LLMs for hate speech detection: Strengths and vulnerabilities. *arXiv Preprint*. https://arxiv.org/abs/2310.12860
7. Kumar, D., AbuHashem, Y. A., & Durumeric, Z. (2024). Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media, 18* (pp. 865–878).
8. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 512–515.

9. Anjum, & Katarya, R. (2024). Hate speech and toxicity detection in online social media: A survey of the state of the art and opportunities. *International Journal of Information Security, 23*(1), 577–608.
10. Faria, F. T. J., Baniata, L. H., & Kang, S. (2024). Investigating the predominance of large language models in low-resource Bangla language over transformer models for hate speech detection: A comparative analysis. *Mathematics, 12*(23), 3687.
11. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759–760).
12. Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information, 15*(9), 517.
13. Akhtar, M. S., Kumar, A., Ekbal, A., & Bhattacharyya, P. (2016). A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016* (pp. 482–493).
14. Syam, S. S., Irawan, B., & Setianingsih, C. (2019). Hate speech detection on Twitter using long short-term memory (LSTM). In *2019 International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 305–310). IEEE.
15. Yang, S., Yu, X., & Zhou, Y. (2020). LSTM and GRU neural network performance comparison study: Yelp dataset case. In *International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)* (pp. 98–101). IEEE.
16. Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of transformer-based models for NLP tasks. In *Proceedings of FedCSIS 2020* (pp. 179–183). IEEE.
17. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).
18. Samghabadi, N. S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., & Solorio, T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 126–131).
19. Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In *International Conference on Data Intelligence and Cognitive Informatics* (pp. 387–402). Springer.
20. Luong, T. S., Le, T.-T., Van, L. N., & Nguyen, T. H. (2024). Realistic evaluation of toxicity in large language models. *arXiv Preprint*. https://arxiv.org/abs/2405.10659
21. Khoma, V., Sabodashko, D., Kolchenko, V., Perepelytsia, P., & Baranowski, M. (2024). Investigation of vulnerabilities in large language models using an automated testing system. In *CEUR Workshop Proceedings, 3826* (pp. 220–228).
22. Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *arXiv Preprint*. https://arxiv.org/abs/1905.12516
23. Hanu, L., & Unitary AI. (2020). *Detoxify* [Software]. GitHub. https://github.com/unitaryai/detoxify
24. Adams, C. J., Borkan, D., Sorensen, J., Dixon, L., Vasserman, L., et al. (2019). *Jigsaw unintended bias in toxicity classification* [Dataset]. Kaggle. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification
25. Martin-ha. (2022). *Toxic-comment-model* [Model]. Hugging Face. https://huggingface.co/martin-ha/toxic-comment-model
26. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint*. https://arxiv.org/abs/1907.11692
27. Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., et al. (2022). ParaDetox: Detoxification with parallel data. In *Proceedings of ACL 2022* (pp. 6804–6818). https://aclanthology.org/2022.acl-long.469
28. Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv Preprint*. https://arxiv.org/abs/2004.06465
29. UC Berkeley D-Lab. (2024). *Measuring hate speech* [Dataset]. Hugging Face. https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech
30. Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv Preprint*. https://arxiv.org/abs/2009.10277
31. Adams, C. J., Sorensen, J., Elliott, J., Dixon, L., McDonald, M., et al. (2017). *Toxic comment classification challenge* [Dataset]. Kaggle. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

Viktor Kolchenko
Postgraduate student, Assistant Professor, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0002-0718-6859
viktor.v.kolchenko@lpnu.ua

Dmytro Sabodashko
PhD, Senior Lecturer, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

Kostiantyn Piataiev
Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0002-9646-7782
kostiantyn.piataiev.kb.2022@lpnu.ua

Vladyslav Horodnyk
Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0005-3169-0870
vladyslav.horodnyk.mkbas.2023@lpnu.ua

Ivan Shchudlo
Postgraduate student, Department of Information and Measurement Technologies
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0003-0412-7682
ivan.o.shchudlo@lpnu.ua

Yuriy Khoma
Dr of Technical Sciences, Associate Professor, Department of Information and Measurement Technologies
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0002-4677-5392
yurii.v.khoma@lpnu.ua

RESEARCH ON MACHINE LEARNING MODELS FOR TOXIC CONTENT DETECTION

Abstract. The rapid growth of digital communication and the proliferation of online platforms have intensified the need to address the problem of detecting toxic content, including hate speech, cyberbullying, threats, and discriminatory statements. Such manifestations negatively affect both individual users and the broader information environment, undermining trust in digital services and contributing to the spread of social bias. Given the impracticality of large-scale manual moderation, automated methods based on Deep Learning and Natural Language Processing (NLP) have become increasingly important.

This study presents a comparative analysis of modern transformer-based models for toxicity detection in text, including Toxic BERT, Toxic Comment Model, RoBERTa Toxicity Classifier, DehaBERT Mono English, as well as the general-purpose large language model Phi-3 mini 4k without task-specific fine-tuning. The evaluation was conducted on two publicly available datasets – Measuring Hate Speech and the Jigsaw Toxic Comment Dataset – using the metrics Accuracy, Precision, Recall, and F1-score. The primary focus was placed on optimizing the F1-score as a balanced measure between precision and recall. Additionally, a threshold-based filtering mechanism was applied to enable objective comparison of models with different sensitivity levels to toxic content.

The experimental results demonstrate that specialized transformer-based models, particularly the RoBERTa Toxicity Classifier, achieve the highest effectiveness in detecting explicit forms of toxic language, reaching accuracy above 90% for categories such as severe hate speech, threats, and

calls for violence. However, performance decreases when addressing context-dependent and indirect manifestations of toxicity, including subtle aggression, disrespect, and offensive humor. It was also shown that class imbalance within the datasets significantly affects classification quality, resulting in lower performance for underrepresented categories. Furthermore, model bias toward specific identity subgroups and sensitivity to profanity were identified.
The study additionally demonstrates that applying a general-purpose LLM without task-specific fine-tuning is both ineffective and computationally inefficient during inference. The obtained results confirm the appropriateness of using specialized models for content moderation tasks and highlight the перспективity of combining general-purpose large language models with domain-specific classifiers to improve context-aware toxicity detection. Future research should focus on developing ensemble approaches and reducing model bias to ensure fairer and more reliable automated moderation systems.

Keywords: toxicity detection; hate speech; Deep Learning; Natural Language Processing; Large Language Models.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Guo, K., et al. (2023). An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)* (pp. 1568–1573). IEEE.
2. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *arXiv Preprint*. https://arxiv.org/abs/2009.11462
3. Biswas, P., & Haritha, D. (2024). Automatic hate speech detection and the hassle of offensive language. *Educational Administration: Theory and Practice, 30*(5), 12663–12668. https://kuey.net/index.php/kuey/article/view/4005
4. Hee, M. S., et al. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 4407–4419).
5. Prem, E., & Krenn, B. (2024). *On algorithmic content moderation*.
6. Roy, S., Harshavardhan, A., Mukherjee, A., & Saha, P. (2023). Probing LLMs for hate speech detection: Strengths and vulnerabilities. *arXiv Preprint*. https://arxiv.org/abs/2310.12860
7. Kumar, D., AbuHashem, Y. A., & Durumeric, Z. (2024). Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media, 18* (pp. 865–878).
8. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 512–515.
9. Anjum, & Katarya, R. (2024). Hate speech and toxicity detection in online social media: A survey of the state of the art and opportunities. *International Journal of Information Security, 23*(1), 577–608.
10. Faria, F. T. J., Baniata, L. H., & Kang, S. (2024). Investigating the predominance of large language models in low-resource Bangla language over transformer models for hate speech detection: A comparative analysis. *Mathematics, 12*(23), 3687.
11. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759–760).
12. Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information, 15*(9), 517.
13. Akhtar, M. S., Kumar, A., Ekbal, A., & Bhattacharyya, P. (2016). A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016* (pp. 482–493).
14. Syam, S. S., Irawan, B., & Setianingsih, C. (2019). Hate speech detection on Twitter using long short-term memory (LSTM). In *2019 International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 305–310). IEEE.
15. Yang, S., Yu, X., & Zhou, Y. (2020). LSTM and GRU neural network performance comparison study: Yelp dataset case. In *International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)* (pp. 98–101). IEEE.
16. Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of transformer-based models for NLP tasks. In *Proceedings of FedCSIS 2020* (pp. 179–183). IEEE.

17. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).
18. Samghabadi, N. S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., & Solorio, T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 126–131).
19. Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In *International Conference on Data Intelligence and Cognitive Informatics* (pp. 387–402). Springer.
20. Luong, T. S., Le, T.-T., Van, L. N., & Nguyen, T. H. (2024). Realistic evaluation of toxicity in large language models. *arXiv Preprint*. https://arxiv.org/abs/2405.10659
21. Khoma, V., Sabodashko, D., Kolchenko, V., Perepelytsia, P., & Baranowski, M. (2024). Investigation of vulnerabilities in large language models using an automated testing system. In *CEUR Workshop Proceedings, 3826* (pp. 220–228).
22. Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *arXiv Preprint*. https://arxiv.org/abs/1905.12516
23. Hanu, L., & Unitary AI. (2020). *Detoxify* [Software]. GitHub. https://github.com/unitaryai/detoxify
24. Adams, C. J., Borkan, D., Sorensen, J., Dixon, L., Vasserman, L., et al. (2019). *Jigsaw unintended bias in toxicity classification* [Dataset]. Kaggle. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification
25. Martin-ha. (2022). *Toxic-comment-model* [Model]. Hugging Face. https://huggingface.co/martin-ha/toxic-comment-model
26. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint*. https://arxiv.org/abs/1907.11692
27. Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., et al. (2022). ParaDetox: Detoxification with parallel data. In *Proceedings of ACL 2022* (pp. 6804–6818). https://aclanthology.org/2022.acl-long.469
28. Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv Preprint*. https://arxiv.org/abs/2004.06465
29. UC Berkeley D-Lab. (2024). *Measuring hate speech* [Dataset]. Hugging Face. https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech
30. Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv Preprint*. https://arxiv.org/abs/2009.10277
31. Adams, C. J., Sorensen, J., Elliott, J., Dixon, L., McDonald, M., et al. (2017). *Toxic comment classification challenge* [Dataset]. Kaggle. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

Отримано редакцією журналу / Received: 16.01.26
Прорецензовано / Revised: 30.01.26
Схвалено до друку / Accepted: 26.03.26