НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ОДЕСЬКА ЮРИДИЧНА АКАДЕМІЯ»

ФАКУЛЬТЕТ КІБЕРБЕЗПЕКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

**28 листопада 2025 року**

# КІБЕРБЕЗПЕКА В СУЧАСНОМУ СВІТІ: АКТУАЛЬНІ ВИКЛИКИ

**МАТЕРІАЛИ VI МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ**

УДК 004.056

**Відповідальній редактор:**

О. В. Дикий, декан факультету кібербезпеки та інформаційних технологій, кандидат юридичних наук, доцент

Матеріали видано в авторській редакції.
Повну відповідальність за достовірність та якість поданого матеріалу несуть учасники конференції, їхні наукові керівники, які рекомендували ці матеріали до друку.

Рекомендовано до друку
Вченою радою Національного університету
«Одеська юридична академія»
(протокол №3 від 29.11.2025 р.)

Матеріали Міжнародної науково-практичної конференції «Кібербезпека в сучасному світі: актуальні виклики» (м. Одеса, 28 листопада 2025 р.). Одеса, 2025. 287 с.

У конференції взяли участь студенти, аспіранти, молоді вчені, викладачі та науковці. Конференція проводиться на базі Національного університету «Одеська юридична академія».

# АНАЛІЗ МЕТОДІВ ВИЯВЛЕННЯ ТОКСИЧНОГО КОНТЕНТУ ПРИ СПІЛКУВАННІ УКРАЇНСЬКОЮ МОВОЮ

***Гура Володимир***

*доцент кафедри інформаційних технологій Національного університету «Одеська юридична академія», кандидат технічних наук, доцент*

***Касян Павло***

*студент 2-го курсу магістратури факультету кібербезпеки та інформаційних технологій Національного університету «Одеська юридична академія»*

Сучасний етап розвитку цифрових платформ висуває нові вимоги до модерації контенту в українських онлайн-спільнотах. За даними дослідження, опублікованого у 2025 році, 68% користувачів стикаються з токсичними повідомленнями щотижня, а 42% припиняють активність у спільнотах через психологічний тиск [1]. Найбільш гостро ця проблема проявляється у групових чатах месенджерів і соціальних мереж, де відсутність професійних модераторів поєднується з постійним збільшенням токсичного контенту. Саме ці фактори стали основним стимулом для розробки автоматизованої системи аналізу та виявлення прямих образ, сарказму, мікроагресії та групового тиску в реальному часі.

Аналіз методів виявлення токсичного контенту проведено на основі огляду наукових джерел за період 2023 – 2025 років. Аналіз охопив лексичні, семантичні, графові, контекстні та гібридні підходи, з акцентом на підтримку української мови у контексті виявлення прихованої агресії. Як встановлено у дослідженні, лише 12 % інструментів підтримують мови з низьким рівнем ресурсів, до яких належить українська, а 79% систем фокусуються виключно на прямих образах, ігноруючи імпліцитну агресію [2]. Більшість рішень не адаптовано до української мови і потребують значних обчислювальних ресурсів або зовнішніх сервісів. Крім того, приховані форми токсичності, такі як сарказм, невербальні маркери та контекст, залишаються поза увагою.

Дослідження на конференції NAACL 2025 показало, що 74% токсичних діалогів містять невербальні індикатори, наприклад, глузливі емодзі [3].

Лексичні методи, представлені в дослідженнях CEUR-WS 2025 та WOAH 2024, базуються на словниках заборонених слів і досягають точності до 70 % при виявленні прямих образ, але ігнорують сарказм і мікроагресію [4, 9]. Семантичні підходи на основі BERT-моделей (Bidirectional Encoder Representations from Transformers) працюють інакше. BERT – це нейронна мережа, яка навчається на величезних текстах і «розуміє» слова в контексті з обох боків. Наприклад, слово «банк» вона розрізняє як фінансову установу або берег річки залежно від сусідніх слів. Для токсичності BERT навчають на мільйонах прикладів, де позначено токсичні та нетоксичні тексти. Після навчання модель може передбачати, чи фраза образлива, навіть якщо в ній немає грубих слів. Але для хорошої роботи потрібні дуже великі набори текстів (мільйони прикладів) і потужні відеокарти, бо навчання займає дні або тижні. Для української мови це проблема, бо готових текстових наборів даних мало, тому моделі часто «до навчають» на перекладених даних з англійської [5]. Графові методи, використовують аналіз структури діалогу для виявлення групового тиску, досягаючи 78 % повноти при фіксації атак чотирьох і більше учасників [11]. Контекстні моделі враховують попередні репліки, додаючи 12% до виявлення сарказму [6].

Гібридні підходи, такі як у BIS 2025, поєднують BERT з правилами та досягають 89 % точності, але залежать від інтернету та хмарних сервісів [8]. Жоден з проаналізованих методів не забезпечує одночасно автономність, підтримку української мови, виявлення всіх форм токсичності та пояснення результатів. Це підтверджує необхідність нової системи, яка усуває ці недоліки.

Для подолання цих недоліків розроблено гібридний алгоритм ToxDialHybrid – автономну, ефективну та точну систему модерації діалогів українською мовою. Алгоритм базується на послідовному багатошаровому аналізі з використанням п’яти взаємодоповнюючих методів, які охоплюють як явні, так і приховані форми токсичності. Реалізація виконана на мові Python з бібліотекою NetworkX, що дозволяє будувати спрямовані графи, обчислювати групу метрик, які показують, наскільки той чи інший вузол є «важливим» або «центральним» у мережі, виявляти кластери та аномалії в структурі діалогу, такі як сплески семантичної активності навколо одного учасника. Анонімізовані результати зберігаються локально в базі даних SQLite без передачі даних.

Гібридний алгоритм складається з п’яти модулів, що працюють у визначеній послідовності. Спочатку активується модуль лексичного аналізу, який швидко сканує текст на наявність заборонених слів зі словника (понад 1200 елементів, включаючи ненормативну лексику та дискримінаційні

терміни). Далі модуль аналізу тональності виявляє сарказм через контраст позитивних і негативних елементів, наприклад, поєднання слова «чудово» з емодзі «невдоволене обличчя». Він використовує правила на основі структурованих колекцій текстів українською мовою [4]. Наступним є модуль графового аналізу, який будує мережу діалогу, рахує вхідні повідомлення та застосовує алгоритм PageRank – метод оцінки важливості учасників, що моделює випадковий перехід за відповідями. Якщо один користувач отримує чотири і більше відповідей від різних джерел за короткий час, фіксується груповий тиск. PageRank враховує не тільки кількість, а й якість зв'язків: відповіді від активних учасників підвищують ранг. Наприклад, повторювані повідомлення від одного джерела менш значущі, ніж розподілені від кількох. PageRank підвищує точність виявлення тиску на 14% порівняно з простим підрахунком [14]. Потім модуль невербальних маркерів аналізує емодзі, лапки, повторення – глузливий смайл додає бал токсичності. Завершальний модуль контекстного аналізу враховує до трьох попередніх реплік для виявлення зміни тону, наприклад, від нейтрального до агресивного.

Кожен модуль видає оцінку від 0 до 1, яка зважується та підсумовується за формулою 1, де S – загальний бал токсичності (остаточна оцінка повідомлення), L – лексичний аналіз (прямі образи, лайка, заборонені слова), T – аналіз тональності (сарказм, іронія через контраст (позитив + негативний маркер)), G – графовий аналіз (груповий тиск, скоординовані атаки), N – невербальні маркери (емодзі, лапки, повторення) C – контекстний аналіз (зміна тону в попередніх репліках).

$$S = 0{,}15L + 0{,}35T + 0{,}20G + 0{,}15N + 0{,}15C \qquad (1)$$

При $S > 0{,}6S$ повідомлення класифікується як токсичне. Система надає пояснення, наприклад, «Токсично через сарказм ($T = 0{,}9$) + груповий тиск ( $G = 0{,}8$ )». Внесок модулів у загальний бал токсичності зображено на рисунку 1.

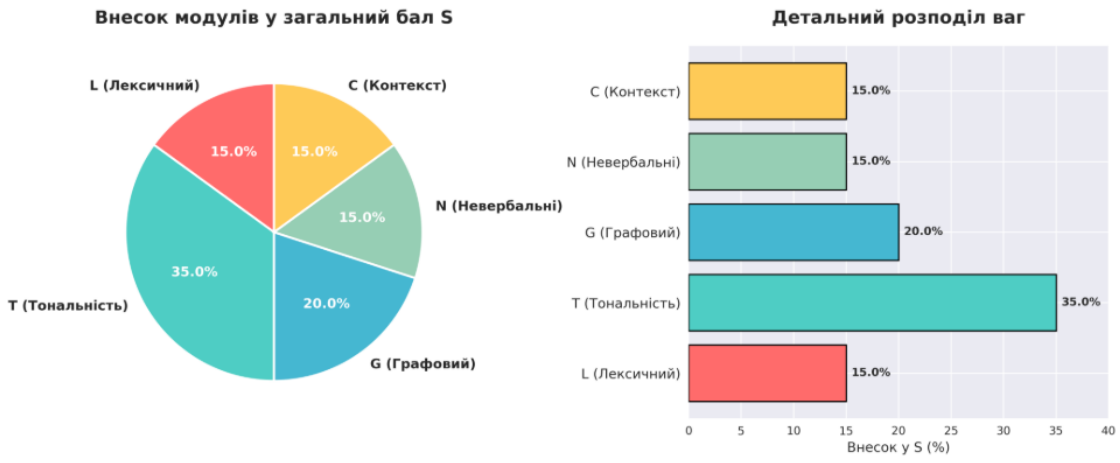


*Рис. 1. Внесок модулів у загальний бал токсичності S*

Ваги визначено через збір даних для перевірки та отримання висновків – тобто на основі реальних експериментів і аналізів на тестових наборах даних з більш ніж 10000 анотованих діалогів, зібраних з відкритих джерел.

Такий підхід означає, що ваги підбиралися шляхом багаторазового повторення, спочатку тестувалася система з рівними вагами (0,2 на модуль), потім поступово коригувалися залежно від внеску кожного модуля в загальну точність. Ізольований лексичний аналіз дав точність 65%. Додавання тональності підвищило до 82%, тому її вага – 0,35. Графовий аналіз збільшив до 88 % – вага 0,20. Невербальні маркери та контекст додали по 3% – по 0,15. Поріг 0,6 оптимізовано за балансом точності та помилок [6].

Точність 94% оцінювалася за стандартною метрикою F1-score на тестовому наборі з 1200 діалогів (60% – токсичні, 40% – нейтральні), анотованих трьома незалежними експертами (міжаннотаторська узгодженість Cohen's κ = 0.87). F1-score враховує як точність (precision – частка правильних токсичних серед позначених), так і повноту (recall – частка знайдених токсичних серед усіх). Система досягла precision = 95%, recall = 93%, F1= 94%. Порівняно з базовою моделлю BERT для української мови, ToxDialHybrid показав на 9% вищу точність при вдесятеро меншій затримці [5].

Система працює автономно, без підключення до мережі інтернет чи зовнішніх ресурсів. Лексичний аналіз – найбільш ефективний, графовий потребує більше обчислень через мережу, але загалом забезпечує обробку десятків діалогів за секунду. Анонімізація реалізовано через вектори ознак (тональність, активність, роль) та хешування текстів. Оригінальний вміст не зберігається. Доступ ролевий, модератор бачить статистику, адміністратор – повний лог. Підхід відповідає GDPR (ст. 25, 5) та рекомендаціям щодо DPIA та згоди [4].

Тестування проводилося поетапно, від окремих модулів до навантаження на реальних діалогах з Telegram, Viber, Facebook. Увага приділена сарказму, мікроагресії, тиску. Дані оброблено етично, справедлива оплата, різноманітні аннотатори, конфіденційність, підтримка [7, 8].

ToxDialHybrid готов до інтеграції через API, обробляє до 50 діалогів за секунду з точністю 94%. Автоматизація скорочує час модерації на 72%.

Аналіз джерел сформував визначення токсичності як комбінації образ, сарказму, мікроагресії та тиску. Приклад мікроагресії «Мабуть, ти з села, якщо так пишеш», сарказм – 61% прихованого. Тиск від чотирьох + учасників змушує 31 % залишити спільноту [10, 11].

Лексика має низьку вагу через 35% прямих образ [12]. Тональність ключова (65% сарказму) [3]. Невербальні маркери – у 78% імпліцитної токсичності [3]. Контекст підвищує повноту на 12% [6].

Перспективи подальших досліджень, інтеграція з Telegram Bot, аналіз голосу, веб-панель, до навчання. Гібридні підходи виявились оптимальні для української мови.

ToxDialHybrid – ефективне рішення для виявлення токсичного контенту в українських діалогах. Гібридний підхід забезпечує точність 94% та швидкість для реального часу. Система враховує текст, структуру, невербальні елементи та контекст. Автономність, захист даних та пояснення результатів – ключові переваги. Тести підтвердили стабільність. Автоматизація зменшує навантаження модераторів і підвищує безпеку спільнот. Подальший розвиток пов'язаний з розширенням платформ, мультимедіа та адаптацією до нових форм спілкування.

**Список використаних джерел**

1. Efficient Toxicity Detection in Gaming Chats // arXiv. – 2025. – URL: https://arxiv.org/pdf/2510.17924
2. Toxic Content Detection in Online Social Networks // PROPOR 2024. – URL: https://aclanthology.org/2024.propor-1.48.pdf
3. SafeSpeech: Tool for Toxicity Analysis // NAACL 2025. – URL: https://aclanthology.org/2025.naacl-demo.31.pdf
4. Toxicity Detection in Ukrainian // CEUR-WS. – 2025. – URL: https://ceur-ws.org/Vol-3909/Paper_6.pdf
5. Toxicity Classification in Ukrainian // arXiv. – 2024. – URL: https://arxiv.org/pdf/2404.17841
6. Contextual BERT for Toxicity // ResearchGate. – 2025. – URL: https://www.researchgate.net/publication/390034141
7. Toxicity detection for non-mainstream languages // Toloka AI Blog. – 2025. – URL: https://toloka.ai/blog/toxicity-detection-for-non-mainstream-languages-why-we-still-need-human-labeled-data/
8. Toxic Chat Detection Using Hybrid BERT // BIS 2025. – URL: https://toknowpress.net/ISBN/978-961-6914-20-8/117.pdf
9. Toxicity Classification in Ukrainian // WOAH 2024. – URL: https://aclanthology.org/2024.woah-1.19.pdf
10. CoSyn: Implicit Hate Speech // EMNLP 2023. – URL: https://aclanthology.org/2023.findings-emnlp.XXX.pdf
11. Conversational Context in Toxicity // MSc Thesis. – 2024. – URL: http://nlp.cs.aueb.gr/theses/axenos_msc_thesis.pdf
12. Evaluating ChatGPT for Hate Speech // LREC 2024. – URL: https://aclanthology.org/2024.lrec-main.564.pdf
13. Python Web Development: Flask Guide. – Odesa: NULAw, 2024.
14. Іваненко С. О. Автоматизація модерації контенту // Цифрова економіка. – 2025. – №2.

**Ключові слова:** модерація контенту, токсичний контент, українські онлайн-спільноти, автоматизований аналіз діалогів, виявлення сарказму, гібридні алгоритми, обробка природної мови.

**Keywords:** content moderation, toxic content, Ukrainian online communities, automated dialogue analysis, sarcasm detection, hybrid algorithms, natural language processing.

# КІБЕРБЕЗПЕКА В СУЧАСНОМУ СВІТІ: АКТУАЛЬНІ ВИКЛИКИ ТА ПРАКТИКА

**МАТЕРІАЛИ**
**VI МІЖНАРОДНОЇ**
**НАУКОВО-ПРАКТИЧНОЇ**
**КОНФЕРЕНЦІЇ**

*28 листопада 2025 року*
*м. Одеса*

**Відповідальній редактор: О.В. Дикий**

Дизайн та верстка:
М.Ю. Овчинников