УДК 004.8

Андрощук 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 НМТіЗ-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.