

АНОТАЦІЯ

Собко Олена Віталіївна. Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії з галузі знань 12 Інформаційні технології за спеціальністю 122 Комп'ютерні науки. – Хмельницький національний університет, Хмельницький, 2025.

Дисертаційна робота присвячена розв'язанню науково-прикладної задачі виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту.

Кіберзалякування стали однією з найбільш поширених форм агресивної поведінки в інтернеті в останні роки. За даними досліджень, приблизно 20–40 % підлітків у всьому світі стають жертвами кіберзалякувань, що негативно впливає на їхнє психічне здоров'я та соціальну взаємодію. Одним з найбільш перспективних підходів до виявлення кіберзалякувань є автоматизований аналіз текстового контенту із застосуванням засобів обробки природної мови. Системи з використанням засобів її обробки вже демонструють високі показники у виявленні кіберзалякувань в текстах соціальних мереж та месенджерів. Однак, незважаючи на те, що автоматизовані системи здатні виявляти кіберзалякування у текстовому контенті, існує низка проблем. Зокрема, проблеми етичної та соціокультурної адаптації алгоритмів та залежність від якісного набору даних впливають на результати аналізу. Крім того, такі моделі часто сприймаються як «чорні скриньки», оскільки їхні результати важко інтерпретувати. Відсутність прозорих механізмів пояснення негативно впливає на їхнє впровадження в системи модерації контенту або правозахисні ініціативи.

Об'єктом дослідження є процес інтелектуального аналізу текстового контенту для виявлення кіберзалякувань.

Предметом дослідження є методи та засоби обробки природної мови для виявлення кіберзалякувань у текстовому контенті.

Метою дослідження є підвищення точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень.

У дисертаційній роботі вперше запропоновано метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечує недискримінацію за віковою, гендерною, релігійною приналежністю, що дозволило підвищити якість навчання класифікаторів для виявлення кіберзалякувань.

У дисертаційній роботі розроблено новий метод виявлення кіберзалякувань у текстовому контенті, який відрізняється від існуючих двоетапним виявленням кіберзалякувань, що полягає у нейромережевій ідентифікації наявності кіберзалякувань та подальшій нейромережевій мультилейбловій класифікації окремих типів кіберзалякувань, що дало можливість підвищити точність та якість виявлення кіберзалякувань.

У дисертаційній роботі також удосконалено метод інтерпретації результатів виявлення кіберзалякувань, який відрізняється від існуючих, можливістю надавати візуальні пояснення для мультилейблової класифікації виявлених типів кіберзалякувань в альтернативних поданнях.

Практичне значення отриманих результатів полягає у доведенні теоретичних результатів дисертаційної роботи та розробці інтелектуальної інформаційної системи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту, що використовує розроблені методи оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, виявлення і класифікації кіберзалякувань, а також інтерпретації результатів виявлення кіберзалякувань, та дозволяє підвищити точність та якість виявлення

кіберзалякувань у текстовому контенті засобами штучного інтелекту й візуально пояснювати прийняті рішення.

Розроблена інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань у текстовому контенті. Інтелектуальна інформаційна система надає можливість оцінювати та коригувати репрезентативність датасетів для навчання моделей машинного навчання за етичними аспектами FATE-принципом справедливості; виявляти та класифікувати типи кіберзалякувань у текстовому контенті. Також інтелектуальна інформаційна система дозволяє отримувати візуальні пояснення для мультитейблової класифікації виявлених типів кіберзалякувань, що сприяє підвищенню довіри до одержаних результатів класифікації типів кіберзалякувань.

Результати дисертаційної роботи впроваджено: у діяльності відділу протидії кіберзлочинам у Хмельницькій області Департаменту кіберполіції Національної поліції України; у ПП «Авіві» (довідка про впровадження); у ГО «ІТ-кластер міста Хмельницького» (довідка про впровадження); у ТОВ «Системи для бізнесу 2» (довідка про впровадження); у навчальному процесі Хмельницького національного університету (акт впровадження); при виконанні держбюджетної теми Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання» (ДР № 0121U112025) (додаток Б).

Ключові слова: кіберзалякування, типи кіберзалякувань, етичні аспекти, FATE-принципи, репрезентативність датасетів, BERT, LIME, інтерпретація результатів, поясненість.

ANNOTATION

Sobko Olena. Methods for detecting and classifying cyberbullying in text content using artificial intelligence. – Manuscript copyright.

Thesis on competition of scientific degree of Doctor of Philosophy by specialty 122 – Computer Science. – Khmelnytskyi National University, Khmelnytskyi, 2025.

The dissertation work is dedicated to solving the scientific and applied problem of detecting and classifying cyberbullying in text content using artificial intelligence.

Cyberbullying has become one of the most common forms of aggressive behavior on the Internet in recent years. According to research, approximately 20–40 % of adolescents worldwide become victims of cyberbullying, which negatively affects their mental health and social interaction. One of the most promising approaches to detecting cyberbullying is automated analysis of text content using natural language processing tools. Systems using natural language processing tools have already demonstrated high performance in detecting cyberbullying in texts on social networks and instant messengers. However, despite the fact that automated systems are able to detect cyberbullying in text content, there are a number of problems. In particular, problems of ethical and sociocultural adaptation of algorithms and dependence on a high-quality dataset affect the results of the analysis. In addition, such models are often perceived as “black boxes” because their results are difficult to interpret. The lack of transparent explanation mechanisms negatively affects their implementation in content moderation systems or human rights initiatives.

Object of the research is the process of intellectual analysis of text content to detect cyberbullying.

Subject of the research is methods and tools of natural language processing to detect cyberbullying in text content.

Purpose of the research is to increase the accuracy and quality of detecting cyberbullying in text content using artificial intelligence with subsequent interpretation of the decisions made.

In dissertation work first proposed a method for assessing and adjusting the representativeness of a dataset based on the FATE principle of fairness, which ensures non-discrimination by age, gender, and religious affiliation, which allowed improving the quality of training classifiers for detecting cyberbullying.

In dissertation work developed a new method for detecting cyberbullying in text content, which differs from existing ones in two-stage detection of cyberbullying, which consists of neural network identification of the presence of cyberbullying and subsequent neural network multi-label classification of individual types of cyberbullying, which made it possible to increase the accuracy and quality of cyberbullying detection.

In dissertation work also improves the method of interpreting the results of cyberbullying detection, which differs from existing ones in the ability to provide visual explanations for multi-label classification of detected types of cyberbullying in alternative representations.

The practical significance of the results obtained lies in proving the theoretical results of thesis and developing an intelligent information system for detecting and classifying cyberbullying in text content using artificial intelligence, which uses the developed methods for assessing and adjusting the representativeness of the dataset according to the FATE principle of fairness, detecting and classifying cyberbullying, as well as interpreting the results of detecting cyberbullying, and allows increasing the accuracy and quality of detecting cyberbullying in text content using artificial intelligence and visually explaining the decisions made.

An intelligent information system has been developed for detecting and classifying cyberbullying in text content. The intelligent information system provides the ability to assess and adjust the representativeness of datasets for training machine

learning models according to ethical aspects of the FATE principle of fairness; to detect and classify types of cyberbullying in text content. The intelligent information system also allows for visual explanations for multi-label classification of detected types of cyberbullying, which helps to increase confidence in the obtained results of classifying types of cyberbullying.

The results of dissertation work were implemented: in Cybercrime Countermeasures Department in Khmelnytskyi Oblast of Cyberpolice Department of Ukraine National Police, in PE «Avivi» (certificate of implementation); in the NGO «IT Cluster of the City of Khmelnytskyi» (certificate of implementation); in LLC «Systems for Business II» (certificate of implementation); in educational process of Khmelnytskyi National University (act of implementation); in the implementation of the state budget theme of Khmelnytskyi National University «Development of information technology for making human-controlled critical safety decisions using mental-formal machine learning models» (DR No. 0121U112025) (Appendix B).

Keywords: cyberbullying, types of cyberbullying, ethical aspects, FATE principles, representativeness of datasets, BERT, LIME, interpretation of results, explainability.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

*Статті у наукових виданнях,
включених до Переліку наукових фахових видань України:*

1. Собко О. В. Нейромережевий пошук і класифікація кіберзалякувань у текстових повідомленнях. *Науковий журнал «Information Technology: Computer Science, Software Engineering and Cybersecurity»*. 2024. № 4. С. 197–205. URL: <https://doi.org/10.32782/IT/2024-4-23>.

2. Собко О. В., Бармак О. В. Метод аналізу та формування репрезентативних вибірок текстових даних із використанням моделей машинного навчання. *Науковий журнал «Computer Science and Applied Mathematics»*. 2024. № 2. С. 83–92. URL: <https://doi.org/10.26661/2786-6254-2024-2-09>.

3. Собко О. В. Метод класифікації кіберзалякувань в україномовному текстовому контенті засобами штучного інтелекту. *Науковий журнал «Наука і техніка сьогодні»*. 2024. № 13 (41). С. 1252–1263. URL: [https://doi.org/10.52058/2786-6025-2024-13\(41\)-1252-1263](https://doi.org/10.52058/2786-6025-2024-13(41)-1252-1263).

4. Собко О. В. Метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту. *Науковий журнал «Вісник Хмельницького національного університету», серія: Технічні науки»*. 2024. № 6, Т. 1 (343). С. 302–309. URL: <https://doi.org/10.31891/2307-5732-2024-343-6-45>.

Публікації, які засвідчують апробацію матеріалів дисертації:

5. Method for Analysis and Formation of Representative Text Datasets / O. Sobko, O. Mazurets, M. Molchanova, I. Krak, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3899, pp. 84–98. URL: <https://ceur-ws.org/Vol-3899/paper9.pdf> (індексована в наукометричній базі Scopus).

6. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, R. Bahrii,

O. Sobko, O. Barmak. *CEUR Workshop Proceedings*, 2024, vol. 3688, pp. 16–28. URL: <https://ceur-ws.org/Vol-3688/paper2.pdf> (індексована в наукометричній базі Scopus).

7. Method for Neural Network Cyberbullying Detection in Text Content With Visual Analytic / I. Krak, O. Sobko, M. Molchanova, I. Tymofiiiev, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3917, pp. 298–309. URL: <https://ceur-ws.org/Vol-3917/paper57.pdf> (індексована в наукометричній базі Scopus).

8. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content / V. Slobodzian, O. Kovalchuk, M. Molchanova, O. Sobko, O. Mazurets, O. Barmak, I. Krak. *CEUR Workshop Proceedings*, 2022, vol. 3171, pp. 561–571. URL: <https://ceur-ws.org/Vol-3171/paper45.pdf> (індексована в наукометричній базі Scopus).

9. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets / O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina. *Lecture Notes on Data Engineering and Communications Technologies*, 2023, vol. 149, pp. 591–607. URL: https://doi.org/10.1007/978-3-031-16203-9_33 (індексована в наукометричній базі Scopus).

Публікації, які додатково відображають наукові результати дисертації:

10. А. с. № 132920 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для оцінювання та коригування репрезентативності текстових датасетів» / О. В. Собко. 2025.

11. А. с. № 132921 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» / О. В. Собко. 2025.