Хмельницький національний університет
Міністерство освіти і науки України

Кваліфікаційна наукова праця
на правах рукопису

СОБКО ОЛЕНА ВІТАЛІЇВНА

УДК 004.8

Д И С Е Р Т А Ц І Я

МЕТОДИ ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

122 Комп'ютерні науки

12 Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

О. Собко Собко Олена Віталіївна

Наукові керівники: Бармак Олександр Володимирович,
доктор технічних наук, професор

Арчіл Чочіа,
доктор філософії,
старший дослідник Школи права
Талліннського технічного
університету (Естонія)

Хмельницький – 2025

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 мультилейблової 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, поясненність.

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.

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. Tymofiiev, 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.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

AI – з англ. artificial intelligence, штучний інтелект;
XAI – з англ. explainable artificial intelligence, пояснюваний штучний інтелект;
NLP – з англ. natural language processing, обробка природної мови;
ML – з англ. machine learning, машинне навчання;
FATE – з англ. Fairness, Accountability, Transparency and Ethics, принципи справедливості, підзвітності, прозорості та етики;
LIME – з англ. Local Interpretable Model-agnostic Explanations, локальні інтерпретовані незалежні від моделі пояснення;
SHAP – з англ. SHapley Additive exPlanations, пояснення Шеплі;
SVM – з англ. support vector machine, метод опорних векторів;
k-NN – з англ. k-nearest neighbors, метод k-найближчих сусідів
LSTM – з англ. long short-term memory, довга короткочасна пам’ять;
GRU – з англ. gated recurrent unit, рекурентний блок із воротами;
BERT – з англ. bidirectional encoder representations from transformers, двонаправлені подання кодувальника на основі трансформерів;
RoBERTa – з англ. robustly optimized BERT approach, покращений BERT;
GPT – з англ. generative pre-trained transformer, генеративний попередньо навчений трансформер;
TF-IDF – з англ. term frequency-inverse document frequency, частотність терміну – обернена частотність документа;
BiLSTM – з англ. bidirectional long short-term memory, двонаправлена довгострокова пам'ять;
SMOTE – з англ. Synthetic Minority Over-sampling Technique, техніка синтетичного збільшення представлення меншості;
ІС – інформаційна система;
TP – з англ. true positive, істинно позитивні випадки;
TN – з англ. true negative, істинно негативні випадки;
FP – з англ. false positive, помилково позитивні випадки;
FN – з англ. false negative, помилково негативні випадки.

ВСТУП

Актуальність теми. Кіберзалякування є однією з найбільш актуальних проблем сучасного онлайн-середовища, яка спричиняє серйозні наслідки для психічного здоров’я користувачів і має значний соціальний вплив. Зі зростанням кількості користувачів соціальних мереж, форумів і месенджерів масштаби цього явища постійно збільшуються, що вимагає розробки засобів для його виявлення та запобігання.

Дослідження показують, що жертви кіберзалякувань частіше стикаються з тривожністю, депресією та зниженням самооцінки, що підкреслює необхідність своєчасного виявлення таких загроз та їхньої нейтралізації [1]. Однак автоматизоване розпізнавання кіберзалякувань у текстах залишається складним завданням через варіативність мовних конструкцій, сарказм, контекстуальну багатозначність та відмінності у сприйнятті агресивності висловлювань залежно від культурних особливостей. Сучасні методи обробки природної мови, зокрема трансформерні моделі, вже демонструють високу ефективність у виявленні агресивного контенту [4]. Проте їхня здатність коректно розпізнавати різні типи кіберзалякувань потребує подальшого вдосконалення.

Крім того, значною проблемою залишається забезпечення етичності роботи таких систем. Автоматизовані системи виявлення кіберзалякувань можуть бути дискримінаційними, тобто такими, коли алгоритми несправедливо оцінюють або дискримінують певні групи користувачів через нерепрезентативність навчальних даних для моделей, які здатні виявляти та класифікувати кіберзалякування.

Також постає проблема прозорості та пояснюваності роботи таких систем, адже користувачі повинні розуміти, які саме висловлювання в текстовому контенті оцінюються як ті, що містять кіберзалякування. Відсутність пояснюваності рішень може викликати недовіру до систем модерації. Через це все більшої уваги набуває розробка пояснюваних моделей штучного інтелекту, які дозволяють користувачам

та модераторам розуміти, які саме мовні ознаки привели до класифікації повідомлення як кіберзалякування.

Таким чином, існує потреба розробки етично відповідальних, пояснювальних методів виявлення та класифікації кіберзалякувань. Використання репрезентативних датасетів для навчання моделей машинного навчання і поясненого штучного інтелекту сприятиме створенню надійних систем, здатних виявляти кіберзалякування та їх типи, водночас забезпечуючи етичність та прозорість прийнятих рішень.

У сфері обробки природньої мови та виявлення кіберзалякувань у текстовому контенті значний внесок зробили як вітчизняні, так і зарубіжні науковці, серед яких можна виокремити роботи таких авторів, як: О. Марченко [151, 152], А. Анісімов [152], Ю. Крак [154], О. Бармак [153, 154], О. Стрижак [155, 156, 157], Х. Ліп'яніна-Гончаренко [158, 159, 160], Б. Нараян (Narayan) [161], Ю. Кумар (Kumar) [162], Х. Шютце (Schütze) [163].

Останнім часом із збільшенням випадків кіберзалякувань в соціальних мережах значно посилилася увага до проблеми їх автоматизованого виявлення у текстовому контенті. Використання методів штучного інтелекту, зокрема, глибоких нейронних мереж та трансформерних моделей, дозволяє автоматизувати процеси виявлення кіберзалякувань, що суттєво знижує навантаження на модераторів контенту. Проте, не зважаючи на значний прогрес у цій галузі, існує низка проблем, що обмежують ефективність існуючих рішень, зокрема, щодо забезпечення етичності та поясненості моделей [126].

Таким чином, існує *суперечність* між можливістю точного виявлення кіберзалякувань у текстовому контенті та відсутністю довіри до навчальних даних, які не можуть гарантувати репрезентативність результатів через відсутність перевірки та можливостей приведення навчальних даних до репрезентативного вигляду. Також існує проблема низької точності класифікації, оскільки деякі типи кіберзалякувань мають спільні ознаки, і, як наслідок, для підвищення точності

класифікації кіберзалякувань вимагається навчання нейромережевої моделі виключно даними з високим рівнем кіберзалякувань, але для визначення рівня кіберзалякувань у тексті має використовуватись клас даних без кіберзалякувань. Дотичною проблемою є зниження довіри до результатів нейромережевих рішень з виявлення кіберзалякувань внаслідок їх низької поясненості. Відтак, забезпечення автоматизованих виявлення та класифікації кіберзалякувань з подальшою інтерпретацією прийнятих рішень є *актуальною науково-прикладною задачею*, шляхом до розв'язання якої пропонується розробка методів і засобів виявлення та класифікації кіберзалякувань у текстовому контенті, що сприятиме підвищенню точності та якості виявлення кіберзалякувань у текстовому контенті з подальшою інтерпретацією прийнятих рішень.

Зазначена науково-прикладна задача відповідає предметній області Стандарту вищої освіти України зі спеціальності 122 – Комп'ютерні науки для третього (освітньо-наукового) рівня вищої освіти, зокрема, такому об'єкту вивчення та діяльності, як «процеси обробки інформації у комп'ютерних системах».

Зв'язок роботи з науковими програмами, планами, темами. Дослідження, результати яких викладено в дисертації, виконано під час виконання окремих розділів науково-дослідної роботи за держбюджетною темою Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання» (ДР № 0121U112025), в яких автор була виконавцем та розробляла нейромережеві архітектури моделей глибокого навчання і виконувала їх опис, які були використані у дисертаційному дослідженні.

Мета і задачі дослідження. *Об'єкт дослідження* – процес інтелектуального аналізу текстового контенту для виявлення кіберзалякувань.

Предмет дослідження – методи та засоби обробки природної мови для виявлення кіберзалякувань у текстовому контенті.

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

Для досягнення поставленої мети необхідно вирішити такі завдання.

1. Провести аналіз методів, засобів та технологій для автоматизованого виявлення кіберзалякувань у текстовому контенті.

2. Розробити новий метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечуватиме недискримінацію за віковою, гендерно та релігійною приналежністю.

3. Розробити новий метод виявлення кіберзалякувань у текстовому контенті.

4. Удосконалити метод інтерпретації результатів виявлення кіберзалякувань.

5. Створити інтелектуальну інформаційну систему для валідації розроблених методів та провести експерименти і порівняння.

Методи дослідження. Для розв’язання поставлених задач використовуються методи математичної статистики та теорії ймовірності, методи машинного навчання, методи чисельної оптимізації, методи обробки природної мови, моделі глибокого навчання.

Наукова новизна дисертаційного дослідження полягає в одержанні таких наукових результатів:

1) 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;

2) 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;

3) 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 і 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-принципу справедливості; виявляти та класифікувати типи кіберзалякувань у текстовому контенті. Також інтелектуальна інформаційна система дозволяє отримувати візуальні пояснення для мультилейблової класифікації виявлених типів кіберзалякувань, що сприяє підвищенню довіри до одержаних результатів їх класифікації.

Результати дисертаційної роботи впроваджено: у відділі 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).

Особистий внесок здобувача та внесок інших співавторів у спільних публікаціях. Список опублікованих праць за темою дисертації наведено в списку використаних джерел – [1–11]. Усі наукові результати дисертаційного дослідження отримані автором особисто. У спільних публікаціях автору належать такі результати: методологія аналізу та формування репрезентативного датасету [5]; розробка методу виявлення образливих висловлювань для української мови; огляд відомих методів та рішень [6]; метод виявлення та класифікації кіберзалякування у текстовому контенті; розробка програмного забезпечення для апробації методу [7]; огляд існуючих підходів до розв’язання задачі; аналіз та визначення доцільних методів пошуку ключових слів у тексті для моделі української мови в завданнях для аналізу спілкування в Інтернеті [8]; аналіз існуючих класифікаторів для бінарної класифікації текстів за сентиментом і обґрунтування доцільності використання конкретних моделей для подальшого проведення дослідження [9].

Внесок інших співавторів у спільних публікаціях: у статті [2] О. Бармак виконував концептуалізацію та складав план дослідження, займався підготовкою чернетки рукопису; у статті [5] О. Мазурець працював над візуалізацією отриманих результатів дослідження, підготовкою чернетки рукопису, виконував адміністрування проєкту; М. Молчанова описувала отримані результати дослідження, виконувала підготовку чернетки рукопису, розробляла програмне

забезпечення для апробації запропонованого методу; Ю. Крак виконував концептуалізацію дослідження, складав план дослідження описаного методу; О. Бармак виконував керівництво проєктом та рецензування чернетки дослідження; у статті [6] Ю. Крак спільно з О. Бармаком виконував концептуалізацію дослідження, керівництво проектом; О. Мазурець виконував підготовку чернетки рукопису, адміністрування проєкту; М. Молчанова складала план дослідження описаного методу, описувала отримані результати дослідження; О. Залуцька розробляла програмне забезпечення для апробації запропонованого методу, Р. Багрій виконував візуалізацію результатів дослідження та підготовку чернетки рукопису; у статті [7] Ю. Крак виконував концептуалізацію дослідження, керівництво проєктом; М. Молчанова виконувала огляд відомих методів і рішень, описувала отримані результати дослідження; О. Мазурець виконував підготовку чернетки рукопису, адміністрування проєкту; І. Тимофієв виконував дослідження на розробленому програмному забезпеченні та описував результати; О. Бармак складав план дослідження, виконував підготовку чернетки рукопису; у статті [8] Ю. Крак спільно з О. Бармаком виконували концептуалізацію дослідження та рецензування рукопису; О. Мазурець керував постановкою експериментів та виконував підготовку чернетки рукопису, адміністрування проєкту; О. Ковальчук спільно з В. Слободзян відповідали за розробку програмного забезпечення для апробації запропонованого методу; М. Молчанова працювала над методологією дослідження, а також опрацьовувала результати дослідження та виконувала підготовку чернетки рукопису; у статті [9] Ю. Крак виконував концептуалізацію дослідження; О. Бармак займався керівництвом проєкту, обговоренням результатів дослідження, рецензуванням та коригуванням рукопису; О. Мазурець працював над методологією дослідження та опрацьовував результати експериментів, виконував підготовку чернетки рукопису, адміністрування проєкту; О. Молчанова виконувала огляд відомих методів та рішень, працювала над методологією дослідження та опрацьовувала результати експериментів;

О. Ковальчук спільно з В. Слободзян розробляли програмне забезпечення для апробації запропонованого методу та проводили експерименти; Н. Савіна виконувала підготовку чернетки.

Апробація матеріалів дисертації. Основні результати дисертаційного дослідження доповідались та обговорювались на міжнародних науково-практичних конференціях та семінарах: 6th International Conference on Computational Linguistics and Intelligent Systems «CoLInS 2022» (12–13 May, 2022, Gliwice, Poland); 16th International Scientific Conference «Intellectual Systems of Decision-Making and Problems of Computational Intelligence ISDMCI-2022)» (June 14–16, 2022, Rivne, Ukraine); Intelligent Systems Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems «ISW-CoLInS 2024» (April 12–13, 2024, Lviv, Ukraine); 1st International Workshop at Advanced Applied Information Technologies «AdvAIT 2024» (December 5, 2024, Khmelnytskyi, Ukraine, Zilina, Slovakia); 7th Workshop for Young Scientists in Computer Science & Software Engineering «CS&SE@SW 2024» (December 27, 2024, Kryvyi Rih, Ukraine).

Публікації. Основні результати дисертації опубліковані у 11 наукових працях ([1–11], додаток А), серед яких 4 статті – у фахових наукових журналах України, включених на дату опублікування до переліку наукових фахових видань України категорії Б; 5 публікацій, які засвідчують апробацію матеріалів дисертації (статті у матеріалах конференцій, що індексуються в наукометричній базі Scopus); 2 авторських свідоцтва.

Структура та обсяг дисертації. Дисертаційна робота складається з анотації, змісту, переліку умовних скорочень, вступу, чотирьох розділів, висновків, списку використаних джерел із 162 найменувань на 22 сторінках і 4 додатків. Загальний обсяг дисертаційної роботи становить 174 сторінки друкованого тексту, із них 137 сторінок основного тексту. Дисертація містить 45 рисунків та 11 таблиць.

РОЗДІЛ 1.
АНАЛІЗ МЕТОДІВ, ЗАСОБІВ ТА ТЕХНОЛОГІЙ ДЛЯ АВТОМАТИЗОВАНОГО ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ

У розділі проаналізовано актуальність застосування засобів і методів штучного інтелекту до виявлення та класифікації кіберзалякувань у текстовому контенті. Виконано огляд відомих підходів аналізу та формування репрезентативних вибірок текстових даних. Проаналізовано відомі системи виявлення та класифікації кіберзалякувань у текстовому контенті, виявлено складнощі класифікації кіберзалякувань. Також проведено огляд відомих систем інтерпретації результатів виявлення кіберзалякувань. На основі проведеного аналізу методів, засобів і технологій для автоматизованого виявлення кіберзалякувань у текстовому контенті зроблено висновки та поставлено задачі з розроблення методів виявлення й класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту.

1.1. Актуальність виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту

На сьогодні залякування є поширеним явищем у суспільстві, що проявляється у систематичному цькуванні однієї особи іншими та супроводжується психологічним, фізичним, економічним чи сексуальним насильством. Термін «залякування» також має синоніми «цькування», «булінг». Залякування особливо небезпечне у закладах освіти, де негативно впливає на емоційний стан, соціальну адаптацію та успішність дітей і підлітків. З метою протидії цьому явищу в Україні запроваджено нормативно-правове регулювання,

що передбачає адміністративну відповідальність за залякування та механізми його запобігання.

Основним документом є Закон України «Про освіту», який визначає залякування як насильницькі дії, в тому числі із застосуванням електронних комунікацій. Відповідно до Закону України від 18 грудня 2018 року № 2657-VIII «Про внесення змін до деяких законодавчих актів України щодо протидії булінгу», залякування визнається адміністративним правопорушенням, за що передбачена відповідальність згідно зі статтею 173-4 Кодексу України про адміністративні правопорушення [13]. Також у 2020 році Міністерством освіти і науки України ухвалено наказ щодо впровадження механізмів запобігання та протидії залякуванням у закладах освіти, що передбачає обов’язкову наявність внутрішніх процедур для реагування на випадки залякувань та проведення роз’яснювальної роботи серед учасників освітнього процесу [14].

Кіберзалякування є формою залякування та часто зустрічається у онлайн-середовищі, насамперед, через свою глобальність, анонімність та швидкість поширення. Цифрові платформи, соціальні мережі, форуми та месенджери відкрили нові можливості для агресивної поведінки, яка має серйозні негативні наслідки для психічного та фізичного здоров'я людей, зокрема дітей і підлітків [4, 15]. Відмінність кіберзалякування від звичайного залякування полягає в тому, що воно здійснюється через засоби електронної комунікації і робить його менш контрольованим та таким, що складно відстежується, адже воно може відбуватися у будь-який час і в будь-якому місці, де є доступ до інтернету [16, 17].

Кіберзалякування проявляється у формі образливих коментарів, публікацій фотографій або відео без згоди людини, поширення чуток та пліток, а також у психологічному тиску, який може здійснюватися через різні цифрові канали. Джерелами поширення кіберзалякувань є переважно соціальні мережі, такі як Facebook, Instagram, TikTok, а також онлайн-форуми, блоги та платформи для обміну відео та фотографіями [18, 19].

Кіберзалякування є глобальною проблемою, що має значний вплив на психічне здоров'я молоді та дорослих. За даними ВООЗ, близько 15 % підлітків у всьому світу зазнають кіберзалякувань. Особливо вразливими є підлітки у віці від 12 до 17 років, де 37 % з них стали жертвами кіберзалякувань. Дослідження показують, що більше 60 % молодих людей стали свідками такого виду насильства, але часто не втручались [20].

В Європі тенденція кіберзалякування також зростає. Останнє дослідження, проведене ВООЗ, показало, що рівень кіберзалякувань серед підлітків зріс з 2018 року здебільшого серед дівчат у віці 11 та 13 років. Зокрема, такі країни, як Литва і Румунія відзначають високі показники серед підлітків, які займаються кіберзалякуваннями [21, 22].

В Україні ситуація з кіберзалякуваннями також є тривожною. За даними опитувань, більше 30 % підлітків стали жертвами онлайн-залякувань, що включають приниження та погрози. Найбільше кіберзалякування спостерігається в соціальних мережах і на платформах для онлайн-спілкування [23].

Кіберзалякування є складним соціальним явищем, що включає різноманітні типи дискримінації, спрямовані на конкретні соціальні групи, зокрема за гендерними, віковими, релігійними та етнічними ознаками [24]. Така класифікація дозволяє більш точно виявляти та аналізувати прояви кіберзалякування [107].

Гендерна дискримінація в контексті кіберзалякувань часто проявляється у вигляді домагань, образ, погроз або принижень. Жінки, дівчата та ЛГБТ-спільноти найчастіше є жертвами таких атак, адже стереотипи про «норму» поведінки та роль жінок у суспільстві стають підґрунтям для вербальних і фізичних атак в мережі. Чоловіки також можуть бути піддані гендерному насильству, але ці випадки зазвичай менш поширені [25].

Щодо вікової дискримінації, кіберзалякування за віковою ознакою часто торкається дітей та підлітків. Молодь, яка активно користується соціальними мережами, є особливо вразливою до образ і знущань [122]. Водночас, старші люди

також стають жертвами кіберзалякувань, оскільки можуть не володіти достатніми навичками для захисту своїх персональних даних чи розуміння загроз, що виникають у цифровому середовищі [26].

Релігійне та етнічне кіберзалякування є одними з найбільш серйозних його проявів. Віруючі, представники меншин або етнічних груп часто стають об'єктами агресії через свої релігійні переконання або етнічну приналежність. У цих випадках напади можуть бути не лише образливими, але й закликати до насильства чи навіть переслідувань, що особливо стосується національних меншин, які можуть бути піддані насмішкам, погрозам або публічному приниженню [27].

Загалом, класифікація кіберзалякувань за цими ознаками допомагає не тільки виявляти конкретні типи агресії, але й розробляти більш ефективні стратегії для протидії цьому явищу [1].

Незважаючи на те, що на соціальних платформах активно впроваджуються інструменти з попередження та виявлення кіберзалякувань, це явище продовжує залишатися широко поширеним. Основна проблема полягає в тому, що такі інструменти здатні працювати лише з інформацією, яка має публічний доступ, тобто з дописами, коментарями та іншими відкритими матеріалами. Водночас, особисте листування користувачів, яке часто стає простором для кіберзалякувань, залишається поза сферою дії цих інструментів, оскільки захищене параметрами конфіденційності соціальних платформ [28].

Існує кілька способів протидії кіберзалякуванням, серед яких важливу роль відіграють технологічні рішення для автоматичного виявлення образливого контенту, правові інструменти, що регулюють онлайн-комунікацію, а також освітні програми, спрямовані на підвищення цифрової грамотності та відповідальності користувачів [29]. Сьогодні в Україні для боротьби з кіберзалякуваннями створено низку ініціатив, спрямованих на забезпечення безпеки у цифровому просторі [123]. Одним з важливих інструментів є «гарячі»

лінії, на які можуть звертатися постраждалі або свідки кіберзалякувань, що особливо важливо для дітей і підлітків, які можуть не знати, як правильно діяти в ситуаціях кіберзалякувань [30]. Також в Україні активно проводяться інформаційно-просвітницькі заходи для молоді, спрямовані на підвищення обізнаності щодо кіберзалякувань. Крім того, існують спеціалізовані вебсайти, на яких можна анонімно повідомити про такі випадки [31].

Таким чином, існуюча ситуація вимагає не лише посилення законодавства, але й розвитку технологій для своєчасного виявлення та боротьби з кіберзалякуваннями [32]. Впровадження методів штучного інтелекту для автоматизації виявлення таких випадків може значно підвищити ефективність боротьби з цією проблемою [33].

Традиційні методи модерації контенту, що базуються на ручному аналізі тексту, не здатні ефективно справлятися з великими обсягами даних та складністю вираження агресії у цифровому форматі [6]. У цьому контексті засоби та методи штучного інтелекту набувають особливої актуальності. Моделі машинного навчання, нейронні мережі автоматизують процес виявлення та класифікації кіберзалякувань, забезпечуючи високу точність і швидкість обробки великих обсягів текстової інформації [34].

Особливу роль у вирішенні цієї проблеми відіграють методи глибокого навчання, які дозволяють не лише виявляти наявність кіберзалякувань у текстах, але й класифікувати їх за різними типами, враховуючи контекстуальні та соціокультурні особливості [35, 36]. Окрім цього, використання ШІ забезпечує можливість реалізації етично відповідальних рішень, що задовольняють вимоги щодо недискримінації та прозорості алгоритмів [5]. Таким чином, впровадження таких методів у сферу виявлення кіберзалякувань є не лише необхідним, а й перспективним напрямом для створення безпечного цифрового простору.

Враховуючи актуальність проблеми кіберзалякування та його негативний вплив на суспільство [37, 38], виникає необхідність у застосуванні нових методів

для розробки систем автоматизованого виявлення та класифікації кіберзалякувань у текстовому контенті. Їх використання дозволить не лише швидко і точно виявляти кіберзалякування, але й аналізувати їх за різними критеріями, зокрема гендерними, віковими, етнічними та релігійними ознаками, що дасть змогу створити більш ефективні стратегії для боротьби з онлайн-залякуванням, підвищуючи рівень безпеки та підтримки вразливих груп користувачів [39].

У сучасному світі активно розробляються численні рішення з використанням штучного інтелекту, покликані вирішувати різноманітні завдання, з якими люди стикаються щодня. Відповідно, результати, що генеруються штучним інтелектом, залежать від навчальних даних (датасетів), на яких вони навчалися, іншими словами, вміст цих навчальних даних безпосередньо впливає на кінцевий результат [2]. Проте кінцеві користувачі часто не знають про вміст навчальних даних, що є значною проблемою. Відсутність прозорості щодо джерел і характеристик даних, які використовуються для навчання алгоритмів штучного інтелекту, підриває довіру до отриманих результатів [100]. В такому випадку часто користувачі не можуть оцінити потенційні упередження чи дискримінаційні елементи, вбудовані у ці алгоритми. Недостатня інформованість про вміст навчальних датасетів збільшує ризик поширення несправедливих або неточних рішень, які можуть мати серйозні наслідки для окремих осіб та суспільства в цілому [2].

На сьогодні текстові датасети створюються для досягнення певної кінцевої мети. Іншими словами, вони репрезентативні відносно мети, для якої вони створені. Репрезентативність – це властивість вибірки, даних або результатів дослідження, яка відображає їхню відповідність і здатність представляти всю сукупність, з якої вони були взяті [40]. Вона означає, що отримані дані адекватно відображають основні характеристики та розподіл ознак генеральної сукупності, забезпечуючи коректність узагальнень і висновків.

У контексті машинного навчання та роботи з даними репрезентативність датасету означає, що він містить збалансовану і різноманітну вибірку прикладів, які відображають реальні умови використання моделі. Це особливо важливо для забезпечення справедливості та уникнення упередженості в моделях машинного навчання [41].

Часто датасети є нерепрезентативними за етичними аспектами, тобто такі датасети не відображають етичну збалансованість і можуть містити дискримінацію щодо певних соціальних груп. Наприклад, якщо модель машинного навчання тренується на даних, у яких переважають зразки з одного соціального чи культурного середовища, вона може не враховувати інтереси та особливості інших груп. Як наслідок, прийняті на основі цих даних рішення можуть бути несправедливими або дискримінаційними. Наприклад, датасети з текстами, які містять, або не містять кіберзалякування. Кількість текстів, які містять кіберзалякування, і кількість текстів, які не містять кіберзалякування, для валідного датасета повинна відповідати генеральній сукупності всіх існуючих текстів з кіберзалякуваннями і без них [7]. Проте зазначений датасет при цьому може не відповідати генеральній сукупності, наприклад, за віковим поділом, що викликає недовіру до рішень, отриманих за моделями, які навчались на цьому датасеті, у представників різних вікових груп тощо [108].

Засоби для оцінювання репрезентативності текстового набору даних відповідно до принципів етичної недискримінації відсутні, хоча це особливо актуально для соціально важливих і чутливих задач, до прикладу виявлення кіберзалякувань, визначення емоційного стану людей за текстовими дописами тощо [40]. Такого роду проблема приводить до того, що отримані результати можуть бути потенційно індивідуально дискримінаційними за різними ознаками, наприклад, такими, як вік, стать, раса, релігія тощо [2]. Відсутність уваги до етичних компонентів при створенні та використанні датасетів приводить до

упередженості в алгоритмах, що негативно впливає на справедливість і достовірність прийнятих рішень [41].

Відомі датасети для навчання нейромереж, наприклад, [42–45] активно використовуються дослідниками, адже мають великий обсяг даних, проте вони не валідувались авторами щодо репрезентативності за принципом справедливості, а, отже, використання таких датасетів для навчання алгоритмів штучного інтелекту можуть потенційно порушувати етичні принципи та, звідси, мати низьку достовірність прийнятих рішень.

Репрезентативність даних у датасетах є однією з вимог для забезпечення надійності наукових досліджень та ефективної роботи моделей машинного навчання [5]. Коли вибірка даних адекватно відображає основні характеристики і структуру всієї популяції, на основі якої вона побудована, це дозволяє отримувати результати, що можуть бути застосовані до загальної сукупності [34]. Таким чином, репрезентативний датасет є запорукою коректності узагальнень і висновків, які базуються на аналізі цих даних [47]. Наприклад, якщо датасет не включає достатньо зразків певної демографічної групи або надмірно зосереджений на інших групах, модель може бути побудована із систематичними похибками, що, у свою чергу, негативно вплине на прогнози та інтерпретацію результатів дослідження [36].

Репрезентативність даних у датасетах не лише впливає на точність результатів і моделей, але й тісно пов'язана з принципами у використанні даних і розробці технологій штучного інтелекту. Якщо датасет не включає належного представлення всіх соціальних, демографічних або культурних груп, це може призвести до дискримінаційних моделей, які надають пріоритет одній групі над іншою, тобто не є справедливими [49, 50]. Забезпечення репрезентативності є важливим для того, щоб моделі були справедливими щодо всіх демографічних груп і уникали системних упереджень, що можуть дискримінувати окремі групи населення. Репрезентативність датасетів за етичним принципом FATE досягається

шляхом коректного балансування за різними етичними аспектами: етнічного, гендерного, релігійного, вікового тощо [51]. Зокрема, принцип справедливості FATE вимагає, щоб дані не були такими, що приводять до дискримінації певних груп. Це означає, що датасет має забезпечувати рівномірне представлення різних соціальних груп, щоби моделі штучного інтелекту працювали однаково точно для всіх груп, не надаючи переваги одній над іншою.

Також у контексті виявлення кіберзалякувань важливим є не лише точність виявлення кіберзалякувань та їх класифікація, але й здатність моделі надавати пояснення своїх рішень. Пояснення рішень, які приймає модель штучного інтелекту, є важливими, оскільки забезпечує розуміння того, чому певний текст був класифікований як кіберзалякування [52, 53].

Алгоритми штучного інтелекту є непрозорими для кінцевого користувача, тобто мають місце ситуації, в яких система не може запропонувати жодної причини чи навести пояснення, на основі якого було прийнято рішення, зазвичай таку проблему називають «проблемою чорної скриньки» [54]. Відсутність прозорості щодо джерел і характеристик даних, які використовуються для навчання алгоритмів штучного інтелекту, підриває довіру до отриманих результатів [55, 56].

Основним завданням інтерпретації результатів є надання чіткої інформації про те, які саме характеристики тексту привели до класифікації як кіберзалякування. Наприклад, модель може виявити негативні елементи, пов'язані з гендером, віком чи етнічним походженням, але для забезпечення довіри до результатів має бути зрозумілим, на яких словах чи фразах базується таке рішення [17, 57]. Принципи інтерпретації включають декілька підходів, які дають можливість з'ясувати, як працює модель, наприклад, через візуалізацію важливих характеристик тексту (наприклад, ключових слів або фраз), які найбільше впливають на прийняття рішення [58, 78]. Також застосовуються методи, що

дозволяють з'ясувати, наскільки сильно ті чи інші частини тексту сприяють класифікації як кіберзалякування.

Зрозумілі та прозорі пояснення необхідні не лише для підвищення довіри до системи, але й для поліпшення корекції та навчання моделей. Якщо система дає пояснення, то користувач може зрозуміти, чому система прийняла саме таке рішення, і, за необхідності, внести зміни в дані або налаштування моделі [79]. Пояснення результатів дозволяють також краще оцінити, чи не відбулося непередбачених упереджень або помилок у роботі алгоритму [80, 81], що в кінцевому підсумку підвищує ефективність таких систем у боротьбі з кіберзалякуваннями [82].

Зважаючи на це, для забезпечення етичності, справедливості та поясненності важливо впроваджувати практики прозорості та підзвітності на всіх етапах створення і використання систем штучного інтелекту. Таким чином, важливо розробляти методи, які будуть забезпечувати репрезентативність даних за FATE-принципом справедливості, що використовуються для навчання моделей машинного навчання. А доручення важливих рішень моделі «чорної скриньки» створює необхідну потребу в тому, щоби алгоритми штучного інтелекту були зрозумілими для процесу прийняття рішень, а, отже, пояснюваними.

1.2. Огляд наявних підходів до аналізу та формування репрезентативних текстових датасетів

При роботі з великими обсягами зібраних для формування датасетів текстових даних наявні дослідження зосереджені на підготовці даних, а саме видаленні некоректних даних (тегів, HTML-коду), надто коротких текстових повідомлень, а також забезпеченні рівномірного розподілу даних серед різних класів або категорій [64].

У більшості випадків дослідження в області виявлення та класифікації кіберзалякувань проводяться на нерепрезентативних даних, що може виникати внаслідок впливу різних чинників, таких як вибірковий збір інформації, обмежене охоплення різних типів кіберзалякувань або недостатня кількість прикладів для певних демографічних груп [65].

На сьогодні також відомі датасети для навчання нейромереж з виявлення кіберзалякувань у текстових дописах користувачів. Ці датасети активно використовуються дослідниками, адже мають великий обсяг даних, які розподілені за різними типами кіберзалякувань. Проте відомі датасети не валідувались авторами на репрезентативність за FATE-принципом справедливості, а, отже, використання таких датасетів для навчання нейронних мереж можуть порушувати принципи справедливості та мати низьку достовірність прийнятих рішень. У випадку задач виявлення та класифікації кіберзалякувань, репрезентативність датасетів має особливе значення, оскільки від неї залежить ефективність і точність моделей глибокого навчання. Використання репрезентативних даних дозволяє знизити ризик упередженості, що може виникнути внаслідок недотримання пропорцій соціальних груп у вибірці [65].

Дослідженню репрезентативності датасетів та справедливому і недискримінаційному представленню демографічних груп у них присвячено багато робіт, оскільки поняття репрезентативності, справедливості та недискримінаційності є важливими у створенні етичних і справедливих моделей машинного навчання [67].

Останнім часом автори все більше уваги приділяють питанню репрезентативності даних у датасетах, проте поточний стан свідчить про те, що існуючі датасети мають прогалини в представленні статі та раси, також характер демографічних змінних робить класифікацію складною і непослідовною. Так, у статті [68] розглядається питання репрезентативності даних у наборах, які включають людей з інвалідністю та літніх людей. Основною метою дослідження є аналіз того, наскільки ці демографічні групи представлені в існуючих датасетах,

що використовуються для розробки алгоритмів штучного інтелекту. Автори виявляють значні прогалини в представленні різних груп за гендером, расою та етнічністю, а також розглядають труднощі, пов'язані зі збором даних від цих груп. Автори описують метрики вимірювання репрезентативності, які включають оцінку розподілу даних за віком, гендером та етнічністю, зокрема порівняння з ідеальними пропорціями, які забезпечили б інклюзивність для всіх груп. Автори використовують демографічні метадані з 190 доступних датасетів для аналізу їх репрезентативності. Щодо балансування датасетів, автори рекомендують підвищити репрезентативність шляхом додавання зразків для недостатньо представлених груп, зокрема шляхом збору додаткових даних або застосування методів синтетичних даних для покращення представленості меншин і людей з обмеженими можливостями.

У статті [69] автори розглядають важливу проблему репрезентативності датасетів у контексті машинного навчання, акцентуючи увагу на необхідності точного відображення популяційних даних. Вони підкреслюють, що для забезпечення якості моделей важливо коректно формувати датасети, які мають бути максимально близькими до реальних характеристик популяції. Основною стратегією, яку вони пропонують для досягнення цього, є використання стратифікованих датасетів, що дозволяють зменшити варіативність між підгрупами та точно відобразити пропорції між різними категоріями у популяції.

Автори дослідження [70] висувають упередження, що виникають як через дисбаланс класів у даних, так і через чутливі аспекти, такі як раса та стать. Вони пропонують новий метод – Fair Oversampling, який поєднує популярний метод SMOTE з модифікаціями, що допомагають знизити вплив чутливих ознак. Наведений підхід передбачає створення синтетичних зразків для менш представлених класів та одночасне «розмиття» захищених ознак, щоб модель менше покладалася на них під час ухвалення рішень. Метод FOS показав покращення в плані як точності, так і справедливості порівняно з іншими методами, такими як стандартний SMOTE або методи перерозподілу ваг. FOS

збільшує точність моделі за рахунок балансування класів і зменшує залежність від чутливих ознак, що покращує групову справедливість. Також автори пропонують нову метрику Fair Utility, яка враховує як точність, так і справедливість.

Дослідники IBM розробили інструментарій AI Fairness 360 з відкритим вихідним кодом для оцінки і зменшення дискримінації в моделях машинного навчання [71]. Основною метою інструментарію є виявлення упередженості за такими атрибутами, як раса, стать або вік, а також надання методів для репрезентативного представлення усіх наведених соціальних груп на різних етапах розробки моделей. AIF360 також дозволяє використовувати різні метрики для виявлення дискримінації, наприклад, Disparate Impact та Equal Opportunity Difference, які допомагають аналізувати, наскільки модель упереджена щодо різних груп. Для зменшення упередженості інструментарій надає алгоритми, що працюють на рівні даних (pre-processing), під час навчання моделі (in-processing) або після отримання результатів (post-processing).

У статті [72] висвітлено проблему інтерсекційних упереджень у моделях обробки природної мови, а саме нерепрезентативного та упередженого представлення різних груп людей у текстових датасетах. Автори аналізують, наскільки сучасні моделі NLP, такі як BERT, RoBERTa і GloVe, виявляють упередження щодо різних демографічних груп, і як ці упередження проявляються на різних завданнях. Проблема, яку розглядають автори, полягає у тому, що попередні дослідження здебільшого фокусувалися на одновимірних упередженнях (наприклад, за статтю або расою), а інтерсекційні – коли взаємодіють кілька демографічних ознак (наприклад, старші жінки з низьким рівнем доходу), залишаються менш вивченими. Для вирішення цієї проблеми автори виконують масштабний аналіз продуктивності та справедливості моделей на п'яти датасетах, що включають до п'яти демографічних вимірів (стать, раса, вік, освіта, дохід). Вони використовують різні техніки зменшення упереджень для порівняння моделей і оцінюють їх продуктивність на завданнях класифікації тексту. Результати показали, що хоча наявні методи зменшення упереджень (наприклад,

для BERT або RoBERTa) добре зберігають прогностичну точність моделей, їх здатність зменшувати інтерсекційні упередження обмежена. В інтерсекційних випадках упередження значно посилюються, і на деяких завданнях відхилення від справедливості може бути на 20–50 % більшим порівняно з окремими демографічними групами.

Автори [73] пропонують спеціалізовану модель машинного навчання для виявлення та мінімізації упередженості в текстових даних, зокрема в новинних статтях. Модель побудована на основі трансформерних моделей, таких як BERT і його оптимізовані версії, такі як DistilBERT, RoBERTa, що здатні ефективно працювати з контекстом тексту, що важливо для виявлення упереджених слів і фраз. Автори стверджують, що їх підхід до мінімізації упередженості є ефективним завдяки глибоким моделям та трансформерним архітектурам, які здатні виявляти й коригувати упередження на різних етапах машинного навчання, починаючи від підготовки даних і до вироблення остаточних прогнозів моделі.

У статті [74] автори наводять проблему гендерної упередженості в моделях обробки природної мови, вирішуючи її за допомогою двох основних підходів: статистичного та каузального забезпечення справедливості. Статистична справедливість передбачає еквівалентні результати для всіх захищених груп, а каузальна справедливість вимагає, щоб модель приймала однакові рішення незалежно від гендерної характеристики індивіда. Однак ці підходи часто дають неоднозначні результати при оцінюванні упередженості. Дослідники пропонують нові методи, що поєднують статистичне і каузальне виправлення для зменшення гендерних упереджень у NLP-моделях. Вони застосовують такі техніки, як зміну вхідних даних для створення альтернативних зразків, а також методи зміни розподілу навчальних даних та зміну ваг даних під час навчання моделі для зменшення упереджень. Результати показали, що поєднання цих технік дозволяє значно зменшити упередження у моделях як за статистичними, так і за каузальними метриками.

Стаття [75] присвячена вирішенню проблеми інтерсекційного упередження в прогнозах моделей машинного навчання, зокрема глибоких нейронних мереж. Інтерсекційне упередження стосується підгруп людей, які мають більше ніж одну захищену характеристику, наприклад, темношкірі жінки, що може призвести до несправедливих результатів у таких моделях, що дискримінує певні групи. Дослідники пропонують новий метод під назвою Fairpriori для автоматичного виявлення упереджених підгруп у даних. Він базується на алгоритмі Apriori, що дозволяє ефективно генерувати підгрупи і обчислювати метрики справедливості для кожної з них. У процесі дослідження Fairpriori підтримує кілька метрик справедливості (демографічний паритет, предиктивний паритет тощо) та має інтерпретовані результати, що полегшує виявлення підгруп, які піддаються дискримінації.

У статті [76] автори виявляють та класифікують упередженість в обробці природної мови. Основна модель, яку вони використовують для цих завдань, заснована на трансформерах, таких як BERT, що є стандартом для роботи з текстовими даними завдяки своїй здатності розуміти контекст. Автори досліджують різні способи виявлення упередженості, зокрема виявлення таких соціальних характеристик, як стать, раса, релігія та сексуальна орієнтація. Для навчання своїх моделей вони використовують кілька публічних наборів даних, більшість з яких зосереджені на виявленні мови ненависті. Модель оцінюється на завданнях класифікації, що використовують як бінарну класифікацію (виявлення упереджених або неупереджених текстів), так і багатокласовість (визначення, проти яких груп спрямовані упередження).

У дослідженні [77] розглядається проблема кіберзалякувань, яка є загрозою для людей на основі різних ознак, таких як релігія, вік, етнічна приналежність і стать. Використаний набір даних авторами було змінено з урахуванням етичних міркувань, що забезпечує відповідальний штучний інтелект. Алгоритм Naive Bayes продемонстрував високу точність, повноту та коректність виявлення

кіберзалякувань, а модель Bi-LSTM показала здатність більш тонко виявляти кіберзалякування.

Отже, у наведених дослідженнях показано, що формування репрезентативних, а, відповідно, і недискримінаційних датасетів є актуальним напрямом дослідження. Тому є доцільним проведення аналізу та коригування кількості зразків у класах датасету для його відповідності до цільових пропорцій згідно з етичними аспектами FATE-принципу справедливості. У результаті датасети повинні містити дані, репрезентативні згідно з пропорціями популяцій за певними критеріями принципу справедливості, наприклад, віком, статтю тощо.

1.3. Аналіз підходів до виявлення та класифікації кіберзалякувань у текстовому контенті

Проблема виявлення кіберзалякувань є складною через низку факторів, серед яких найбільш важливими є контекстуальна багатозначність мови, культурні відмінності та лінгвістичні нюанси. Одним з основних труднощів при виявленні кіберзалякувань є необхідність відрізняти агресивну поведінку від інших форм комунікації, таких як сарказм, жарти чи просто дружні кепкування [78]. Ці види комунікації можуть використовувати схожі лексичні та стилістичні засоби, що створює суттєві перешкоди для автоматизованих систем аналізу тексту, навіть при використанні сучасних алгоритмів машинного навчання. Сарказм та жарти часто використовують іронію, гіперболу або інші риторичні прийоми, де поверхневий зміст висловлювання не відповідає справжнім намірам мовця. Наприклад, фраза, що виглядає як образа, насправді може бути жартом між близькими друзями.

У роботі [67] автори порівнюють різні методи на основі BERT для виявлення кіберзалякування. Ними проведено аналіз помилок, який показав, що багато хибно класифікованих текстів є саркастичними. Для зменшення таких помилок автори запропонували використати багатошаровий персептрон, що

виявляє сарказм при класифікації, і це дало кращі результати порівняно з іншими методами.

Іншою складністю є те, що кіберзалякування часто включає приховану агресію або маніпулятивну поведінку, що не завжди очевидна при поверхневому аналізі тексту. Наприклад, агресор може використовувати пасивно-агресивні висловлювання або специфічні натяки, які мають образливий підтекст, але на рівні слів можуть бути абсолютно нейтральними, що ускладнює задачу для систем, які не здатні цілком зрозуміти підтекст або інтенцію повідомлень [80].

У статті [81] розглядаються два найпоширеніші прояви кіберзалякувань – агресія та образлива мова. Автори представляють новий, вручну анотований, набір даних, що включає 10,000 твітів англійською та змішаною хінді-англійською мовами, які були анотовані для виявлення агресії та образливого змісту. Для перевірки узгодженості анотацій між кількома анотаторами було отримано коефіцієнти згоди на рівні 67 % та 74 % для відповідних завдань, що вказує на значну узгодженість. Дослідники провели ретельне доопрацювання попередньо навчених мовних моделей, використовуючи цей набір даних, щоб оцінити його ефективність. Найкращі моделі на тестових наборах досягли макросередніх значень F1-метрик на рівні 67,87 % та 65,45 %, відповідно, для двох завдань.

Крім того, важливу роль відіграють культурні та соціальні фактори, адже у різних культурах однакові висловлювання можуть мати різні конотації. Те, що в одній культурі сприймається як жарт, в іншій може бути сприйняте як серйозна образа. Наприклад, у роботі [82] автори для вирішення особливостей турецької мови використовують інструмент обробки природної мови Zemberek-NLP, який вловлює нюанси мови, підвищуючи точність моделі виявлення. Автори стверджують, що деякі фрази чи терміни, які можуть здаватися нешкідливими чи навіть буденними в певних культурах, можуть мати агресивні конотації в турецькому контексті, що вимагає додаткових інструментів для виявлення кіберзалякувань. Таким чином, використовуючи наведений підхід та

використовуючи класифікатор SVM, досягають точності у виявленні кіберзалякувань в 95,9 %.

На сьогодні існують різні підходи до виявлення кіберзалякувань – це бінарна класифікація, суттю якої є поділ тексту на два класи: «кіберзалякування» або «некіберзалякування». Багатокласовий підхід, який передбачає класифікацію тексту на кілька різних типів кіберзалякувань, що можуть містити більш деталізовані класи, наприклад, кіберзалякування за релігійною ознакою, віковою, гендерною чи етнічною тощо. Також проводять і мультилейблову класифікацію, яка передбачає виявлення одночасно декількох типів кіберзалякування в одному текстовому зразку.

У дослідженні [83] розглядаються методи машинного та глибокого навчання для виявлення кіберзалякувань. Серед протестованих моделей Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN, BERT найкращі результати показала модель BERT, досягнувши точності 88,8 % у бінарній класифікації та 86,6 % – у мультилейбловій.

У дослідженні [84] розглядаються проблеми дисбалансу класів у наборах даних для виявлення кіберзалякувань, що ускладнює роботу алгоритмів машинного навчання, особливо в бінарних наборах. Для вирішення цієї проблеми застосовуються методи підвищення та зменшення вибірок, такий як SMOTE. Результати показують, що ефективність технік балансування залежить від розміру набору даних, рівня дисбалансу та використовуваного класифікатора, причому жоден з методів не демонструє постійної переваги. Автори, застосовуючи методи балансування класів у датасетах, досягли точності у 99 % з використанням класифікатора SVC.

У дослідженні [85] автори пропонують нову теорію для виявлення кіберзалякування. Моделі Support Vector Machine, Naive Bayes і Logistic Regression були протестовані разом з різними методами обробки природної мови. Автори стверджують, що точність виявлення кіберзалякувань покращується за допомогою аналізу настроїв, аналізу N-грамів та інших нетрадиційних методів виділення

ознак, таких як TF-IDF і виявлення ненормативної лексики. За допомогою комбінованого підходу досягають точності у 75,17 %.

Стаття [86] пропонує новий підхід до виявлення образливих та агресивних коментарів за допомогою моделей глибокого навчання, таких як глибокі нейронні мережі, мережі на основі радіальних базисних функцій, які комбінуються з оптимізатором Adam. Автори демонструють, що запропонований підхід значно підвищує точність виявлення кіберзалякувань за допомогою вибору найкращих ознак для аналізу тексту.

Автори [87] зосереджуються на глибокому навчанні як перспективному підході для розв'язання задачі виявлення кіберзалякувань. Вони детально розглядають різні моделі, такі як CNN, RNN, LSTM, GRU та трансформери, порівнюючи їхню ефективність у завданні виявлення кіберзалякувань. Дослідження показує, що рекурентні нейронні мережі (зокрема, LSTM і GRU) демонструють найкращі результати при роботі з послідовними даними.

У роботі [88] проводиться оцінка ефективності різних алгоритмів глибокого навчання (LSTM, GRU, CNN-BLSTM та ін.) на арабських датасетах. Як результат, запропоновано гібридну модель глибокого навчання, яка поєднує найкращі характеристики базових моделей CNN, BLSTM і GRU для виявлення кіберзалякувань. Ця модель підвищує точність класифікації на всіх досліджуваних наборах даних і може бути інтегрована в різні соціальні мережі для автоматичного виявлення кіберзалякувань в арабськомовних текстах.

Автори роботи [89] представили модель глибокого навчання для виявлення кіберзалякувань на основі даних з бенгальських соціальних мереж, що включають 12 282 різноманітних коментарі. Автори використовують двошарову модель Bi-LSTM, застосовуючи різні оптимізатори та 5-кратну кросвалідацію. Результати дослідження показують, що модель досягла точності 94,46 % при використанні оптимізатора SGD (стохастичний градієнтний спуск), а також вищої точності 95,08 % і F1-оцінки 95,23 % при застосуванні оптимізатора Adam. Модель також показала точність 94,31 % при 5-кратній кросвалідації.

Наведені дослідження демонструють, що такі архітектури, як LSTM, GRU, Bi-LSTM та їх гібридні варіанти, покращують точність і ефективність виявлення кіберзалякувань, особливо в контексті обробки послідовних даних, де ці моделі демонструють високі результати. Крім того, попередньо навчені мовні моделі, такі як BERT і XLNet, також виявилися ефективними, особливо при роботі з великими наборами даних, а, отже, є доцільним використання цих моделей для виявлення та класифікації кіберзалякувань у текстовому контенті.

Отже, з огляду на проведений аналіз, можна відзначити, що сучасні підходи до виявлення кіберзалякувань намагаються поєднувати різні методи. До таких підходів належать лінгвістичний аналіз, який враховує семантику та синтаксис тексту, а також контекстуальні моделі, що включають інформацію про попередні повідомлення у розмові або особисті стосунки між комунікантами. Також науковці проводять бінарну, багатокласову та мультилейблову класифікацію для виявлення кіберзалякувань і його типів. Однак ці моделі все ще залишаються недосконалими. Незважаючи на значні досягнення в застосуванні моделей глибокого навчання для виявлення кіберзалякувань, наведені дослідження мають недоліки, особливо у контексті етичних аспектів. Роботи зосереджені на технічних показниках точності та ефективності моделей, тоді як етичні питання, пов’язані з виявленням кіберзалякувань у текстовому контенті, часто залишаються поза увагою. Питання недискримінаційності моделей через використання неоднорідних або недостатньо репрезентативних датасетів для навчання моделей не розглядаються належним чином.

1.4. Огляд наявних систем інтерпретації результатів виявлення кіберзалякувань

Штучний інтелект став невід’ємною частиною використання сучасних технологій, що змушує зосереджувати увагу не лише на продуктивності систем, а й на прозорості алгоритмів та їх здатності надавати зрозумілі обґрунтування своїх

дій. Відповідно до Загального регламенту захисту даних Європейського Союзу, користувачі мають право отримувати пояснення стосовно конфіденційності, захисту даних та роботи алгоритмів, що підкреслює важливість розробки пояснюваних ШІ-систем [90]. Тому пояснювальний штучний інтелект є важливою концепцією у сфері сучасних технологій, що передбачає створення систем штучного інтелекту, здатних надавати зрозуміле пояснення своїх рішень користувачам.

Пояснювальний штучний інтелект може надавати пояснення у двох основних формах: візуальній та текстовій [91]. Візуальні пояснення використовують графічні елементи для відображення результатів моделей машинного і глибокого навчання. Текстові пояснення, в свою чергу, надають аргументацію у вигляді слів, речень або навіть природної мови [92]. Ця форма пояснення широко використовується у рекомендаційних системах, включаючи певні XAI-системи прогнозування [93]. В якості інтерпретаційних моделей для пояснення отриманих результатів часто використовуються методи [132]:

– Local Interpretable Model-agnostic Explanations, який генерує локальні пояснення для кожного передбачення, показуючи, які слова найбільше вплинули на результат;

– SHapley Additive exPlanations, що базується на теорії ігор і обчислює внесок кожного слова у передбачення, враховуючи взаємодію між ознакам;

– Transformers Interpret, що є інтерпретаційною бібліотекою, яка спеціально розроблена для роботи з моделями на основі нейромереж трансформерів, такими як BERT, GPT, RoBERTa та іншими моделями з бібліотеки Hugging Face;

– методи на основі Attention, які дозволяють аналізувати ваги уваги трансформерів (наприклад, у моделі BERT) для розуміння важливості окремих слів чи фраз у прийнятті рішень моделі.

Деякі з авторів також пропонують підходи до інтерпретації результатів виявлення кіберзалякувань у текстовому контенті. Наприклад, у [94] пропонується

уніфікована модель BiLSTM-LIME для класифікації контенту кіберзалякування на платформі Twitter. Автори стверджують, що техніка LIME надає пояснення високого рівня, висвітлюючи найбільш доречні токени, які сприяли прийняттю рішення моделлю.

Автори дослідження [95] запропонували новий підхід до виявлення та класифікації кіберзалякування у текстах соціальних медіа за допомогою ансамблю BERT та SVM з пошуком у сітці для багатокласової класифікації. Порівняння з іншими моделями машинного та глибокого навчання показало, що запропонована модель досягає точності 90 % на тестових даних, перевершуючи інші. Для інтерпретації прогнозів використано техніку SHAP.

Робота [96] присвячена виявленню образливого та дискримінаційного контенту, також відомого як мова ненависті, в текстових даних. У дослідженні порівнюються три техніки векторизації тексту (CountVectorizer, GloVe та BERT) у поєднанні з двома моделями машинного навчання (SVM та логістична регресія). Для пояснення рішень моделі використано методи XAI, зокрема LIME та SHAP, що дозволяють інтерпретувати результати моделі машинного навчання.

У дослідженні [97] автори використовують техніку SHAP для пояснення результатів моделей глибокого навчання. SHAP дозволяє інтерпретувати передбачення моделей, надаючи локальні пояснення для кожного окремого прогнозу, а, отже, техніка показує, які ознаки найбільше вплинули на результат моделі, дозволяючи користувачам зрозуміти, чому певний твіт було класифіковано як образливий.

Отже, хоча в ряді досліджень активно використовуються різні інтерпретаційні моделі для пояснення рішень штучного інтелекту, більшість з них зосереджена на трактуванні результатів для задач бінарної класифікації. Однак у таких дослідженнях не розглядається підхід, який би дозволяв інтерпретувати рішення мультилейблових моделей, що застосовуються для визначення різних типів кіберзалякувань в тексті. Це є важливою проблемою, оскільки для

ефективного виявлення різноманітних проявів кіберзалякувань, таких як гендерні, вікові, етнічні або релігійні, необхідно мати прозорі та зрозумілі алгоритми пояснення результатів роботи складних нейромережевих моделей для мультилейблової класифікації.

1.5. Висновки. Постановка задачі

Аналіз методів, засобів і технологій для автоматизованого виявлення кіберзалякувань у текстовому контенті показав, що, незважаючи на постійний розвиток і вдосконалення комп'ютерних систем, існують проблеми в ефективному виявленні кіберзалякувань в різних його формах. Використання штучного інтелекту, зокрема методів машинного навчання, дозволяє значно підвищити точність виявлення різних типів кіберзалякувань, таких як вікові, гендерні, етнічні та релігійні залякування. Однак, через складність та багатогранність соціальних явищ, досліджені технології вимагають подальшого розвитку, особливо у напрямі досягнення репрезентативності датасетів, на яких навчаються моделі.

Використання нейромережевих архітектур для виявлення та класифікації кіберзалякувань дозволяє досягати високих результатів точності при аналізі текстового контенту. Водночас, необхідність в аналізі та коригуванні репрезентативності датасетів на етапі їх підготовки та навчання моделей забезпечує усунення їх дискримінаційності і підвищує якість результатів. Разом з тим, розробка та інтеграція інструментів для інтерпретації рішень штучного інтелекту дозволяє не тільки поліпшити розуміння результатів, а й підвищити прозорість і довіру до таких технологій, що є особливо важливим в контексті етичних вимог.

Проаналізовані існуючі підходи підтверджують, що ефективність виявлення кіберзалякувань в текстовому контенті залежить не тільки від нейромережевих моделей, але й від їх здатності до надання чітких і зрозумілих пояснень своїх

рішень, що дозволяє підвищити якість виявлення кіберзалякувань, забезпечуючи етичність, прозорість і справедливість використання цих технологій у реальних умовах. Тому підвищення точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень є актуальною науково-прикладною задачею.

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

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.

Для досягнення мети роботи необхідне виконання таких завдань:

1) провести аналіз методів, засобів та технологій для автоматизованого виявлення кіберзалякувань у текстовому контенті;

2) розробити новий метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечуватиме недискримінацію за віковою, гендерною і релігійною приналежністю;

3) розробити новий метод виявлення кіберзалякувань у текстовому контенті;

4) удосконалити метод інтерпретації результатів виявлення кіберзалякувань;

5) створити інтелектуальну інформаційну систему для валідації розроблених методів і провести експерименти та порівняння.

РОЗДІЛ 2.
МЕТОДИ ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

У розділі розглянуто підхід для виявлення та класифікації кіберзалякувань в текстовому контенті засобами штучного інтелекту. Запропоновано три методи, які дозволяють вирішити такі проблеми щодо виявлення і класифікації кіберзалякувань у текстовому контенті:

1. Суперечність між можливістю точного виявлення кіберзалякувань у текстовому контенті і відсутністю довіри до навчальних даних, які не можуть гарантувати репрезентативність результатів через відсутність перевірки та можливостей приведення навчальних даних до репрезентативного вигляду.

2. Низька точність класифікації, оскільки деякі типи кіберзалякувань мають спільні ознаки, і, як наслідок, для підвищення точності класифікації кіберзалякувань вимагається навчання нейромережевої моделі виключно даними з високим рівнем, але для визначення рівня кіберзалякувань у тексті має використовуватись клас даних без кіберзалякувань.

3. Зниження довіри до результатів нейромережевих рішень з виявлення кіберзалякувань внаслідок їх низької пояснюваності.

У розділі наведено обмеження запропонованого підходу, а також описано метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, метод виявлення та класифікації типів кіберзалякувань у текстовому контенті, метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті. Запропонований підхід дозволяє оцінювати та коригувати датасети за такими етичними аспектами: вік; гендер; релігія, а також виявляти та класифікувати кіберзалякування таких типів:

– етнічне;

– релігійне:

– вікове;

– гендерне;

– інші типи.

Для інтерпретації результатів виявлення кіберзалякувань у текстовому контенті пропонується використати метод LIME, що дозволить сформувати такі способи пояснення:

– колірна схема з використанням абсолютного значення ваг з метою визначення яскравості кольору для інтерпретації результатів виявлення різних типів кіберзалякувань;

– колірна схема з відображенням кольору та його яскравості для інтерпретації результатів виявлення типів кіберзалякувань з урахуванням негативного чи позитивного типу впливу на результат;

– діаграми для графічної інтерпретації впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування.

– діаграма з відображенням середнього значення важливості топ 10 слів для всіх класів.

Також запропонований підхід має ряд обмежень, що впливають на якість виявлення кіберзалякувань у текстовому контенті.

2.1. Підхід до виявлення та класифікації типів кіберзалякувань у текстовому контенті

У сучасному соціальному просторі дуже переплетена інформація людського та штучного походження. І вся вона є потенційним носієм дискримінаційного контенту. Проблема автоматизованого виявлення дискримінаційного контенту охоплює різну його природу: свідомо створеного людиною, згенерованого штучним інтелектом на вимогу людини і несвідомо доданого в повідомлення людиною [100].

На сучасному етапі у зв'язку з відсутністю засобів для оцінювання репрезентативності текстових наборів даних відповідно до принципу етичної недискримінації, текстові датасети для навчання штучного інтелекту не збалансовані за етично важливими ознаками. Внаслідок навчання штучного інтелекту з потенційною індивідуальною дискримінацією за різними типами кіберзалякувань, одержуються етично некоректні результати (рис. 2.1).

Рис. 2.1 – Поточний і запропонований підхід до виявлення кіберзалякувань з урахуванням FATE-принципу справедливості

Відтак, пропонується виконувати виявлення і класифікацію кіберзалякувань у текстовому контенті засобами штучного інтелекту з урахуванням FATE-принципу справедливості. Для цього підхід передбачає оцінку та коригування репрезентативності наборів текстових даних відповідно до принципу справедливості – етичної недискримінації за віком, статтю, расою, релігією тощо [101]. Одержаний індивідуально неупереджений збалансований текстовий датасет забезпечує навчання штучного інтелекту за принципом етичної недискримінації за

різними типами кіберзалякувань. Як наслідок, штучний інтелект дає етично коректні результати, що дозволяє підвищити якість навчання класифікаторів для виявлення кіберзалякувань [103].

Запропонований у дисертаційному дослідженні підхід до виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту складається з послідовних етапів. Ці етапи забезпечують виявлення кіберзалякування та класифікацію їх типів з врахуванням FATE-принципу справедливості, а також інтерпретацію наданих нейромережевою моделлю результатів щодо класифікації [104]. Схема запропонованого підходу подана на рис. 2.2.

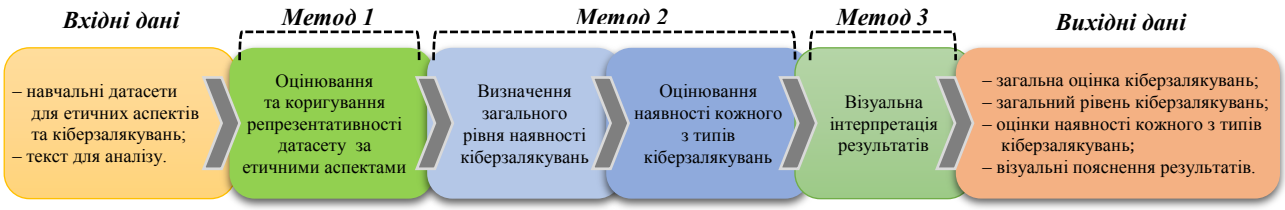


Рис. 2.2 – Підхід до виявлення та класифікації типів кіберзалякувань у текстовому контенті

Вхідними даними є навчальні текстові датасети для оцінювання та коригування репрезентативності за етичними аспектами FATE-принципу справедливості, текстові датасети для виявлення та класифікації типів кіберзалякувань, а також текстовий контент для аналізу. Для подальшої обробки даних запропоновано реалізацію трьох методів, що виконуються послідовно.

Метод 1, наведений в пункті 2.5, дозволяє оцінювати та коригувати за FATE-принципом справедливості репрезентативність датасетів, які призначені для виявлення і класифікації типів кіберзалякувань. В результаті виконання кроків методу отримуються датасети, що є репрезентативними за етичними аспектами. Такі датасети використовуються з метою навчання моделей машинного навчання

для виявлення та класифікації типів кіберзалякувань у вхідному текстовому контенті.

Метод 2, наведений в пункті 2.6, дозволяє виявляти та класифікувати кіберзалякування у текстовому контенті. Наведений метод є двоетапним: на першому етапі дозволяє визначити загальну оцінку наявності та загальний рівень кіберзалякувань у текстовому контенті, на другому – визначити числову оцінку наявності кожного з типів кіберзалякувань, якщо загальний рівень кіберзалякувань у текстовому контенті є високим.

Метод 3, наведений в пункті 2.7, формує множину візуальних подань пояснень результатів класифікації кіберзалякувань, для яких використовуються обраховані оцінки наявності кожного з типів кіберзалякувань у тексті, що досліджується.

Результатом аналізу тексту за запропонованим підходом є вихідні дані, що містять загальну оцінку та загальний рівень кіберзалякувань, оцінки наявності кожного з типів кіберзалякувань і візуальні пояснення результатів. Наведеним чином забезпечується виявлення та класифікація кіберзалякувань у тестовому контенті, що є комплексним підходом, який передбачає поетапне та послідовне виконання запропонованих у дисертаційному дослідженні методів.

Відповідно до рис. 2.2, на першому етапі оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості використовуються датасети, призначені для виявлення і класифікації типів кіберзалякувань. Датасети аналізуються та коригуються відповідно до етичних аспектів, таких як віковий, гендерний, релігійний аспекти. Для цього застосовуються попередньо навчені моделі машинного навчання, здатні аналізувати відповідність датасету зазначеним критеріям для кожного етичного аспекту, що забезпечує формування репрезентативного датасету. Усі ці кроки спрямовані на створення репрезентативних та недискримінаційних датасетів, які

використовуються на другому етапі для навчання моделей, призначених для виявлення та класифікації кіберзалякувань (рис. 2.3).

Рис. 2.3 – Схема підходу до виявлення та класифікації кіберзалякувань у текстовому контенті

На другому етапі, що передбачає виявлення та класифікацію кіберзалякувань у текстовому контенті, використовуються основні компоненти: репрезентативні текстові датасети, створені на попередньому етапі; текстові зразки для аналізу та нейромережеві моделі. Перша з моделей здійснює бінарну класифікацію, визначаючи наявність чи відсутність кіберзалякувань у тексті, тоді як друга класифікує виявлені випадки за їхніми типами. Завдяки цим моделям стає можливим аналіз текстового контенту та точне віднесення повідомлень до відповідних типів кіберзалякувань [103]. На виході цього етапу отримується загальна оцінка рівня кіберзалякувань у тексті, а також деталізована інформація про кожен його тип у досліджуваному текстовому контенті [104].

Третій етап спрямований на інтерпретацію результатів, отриманих під час класифікації. Зокрема, інформація про виявлені типи кіберзалякувань використовуються для візуального подання впливу окремих слів на прийняті рішення нейромережевої моделі щодо мультилейблової класифікації.

Таким чином, запропонований підхід є комплексним і включає всі етапи аналізу текстового контенту для виявлення та класифікації кіберзалякувань. Він враховує етичні аспекти FATE-принципу справедливості, забезпечуючи репрезентативність та недискримінаційність датасетів для навчання моделей, що підвищує якість виявлення кіберзалякувань. Також надає можливість інтерпретації результатів класифікації за допомогою візуального подання впливу окремих слів на прийняті рішення моделі щодо виявлених типів кіберзалякувань.

Відповідно, на кожному етапі реалізовано запропоновані методи: оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості (п. 2.5), виявлення та класифікації типів кіберзалякувань (п. 2.6), інтерпретації результатів виявлення кіберзалякувань у текстовому контенті (п. 2.7). Запропонований підхід має ряд обмежень, що впливають на точність виявлення та класифікації кіберзалякувань у текстовому контенті, які описані у пункті 2.2.

2.2. Обмеження підходу до виявлення кіберзалякувань у текстовому контенті

Розроблені методи (метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, метод виявлення кіберзалякувань у текстовому контенті, метод інтерпретації результатів виявлення кіберзалякувань) мають обмеження, які необхідно врахувати у подальшому дослідженні. Так, запропоновані у дисертаційному дослідженні методи виявлення кіберзалякувань у текстовому контенті працюють з українською мовою, тому вони не можуть бути безпосередньо застосовані до текстів іншими мовами без попередньої адаптації

моделей машинного та глибокого навчання. Для роботи з іншими мовами необхідно використати відповідні датасети мовою, на якій необхідно для мови вхідних текстових повідомлень, які потрібно проаналізувати на наявність кіберзалякувань, навчити моделі машинного і глибокого навчання.

З наведеного випливає ще одне обмеження, а саме наявність таких датасетів, які можуть бути використані для навчання моделей на виявлення і класифікації кіберзалякувань. Через відсутність україномовних датасетів для навчання моделей, було використано автоматизований переклад за допомогою бібліотеки Google Translate. При перекладі текстів з англійської на українську мову виникають труднощі, пов’язані з відсутністю точних відповідників, зміною граматичних конструкцій і тональності, що в подальшому впливає на точність класифікації. Наприклад, англійський вираз «gaslighting» не має точного відповідника в українській мові. Його можна передати як «маніпуляція свідомістю» або «психологічний тиск», але ці варіанти можуть не повністю відображати значення терміна, що може призвести до помилок у визначенні типу кіберзалякувань. Подібна ситуація виникає і з «trigger warning», що іноді перекладають як «застереження про тригер», хоча цей вираз не є природним для української, а це також ускладнює коректне визначення контексту.

Проблема суржику також впливає на якість тексту. Наприклад, англійське «bullying» часто неправильно передають як «булінг» замість коректного «цькування», а «toxic» – як «токсичний» замість «отруйний» (залежно від контексту). Використання таких запозичень у навчальних даних приводить до неоднозначності розпізнавання кіберзалякувань моделлю, оскільки вона навчатиметься на непослідовних мовних конструкціях.

Зміна граматичних конструкцій також впливає на точність класифікації. В англійській мові часто використовуються активні конструкції, тоді як в українській – пасивні. Наприклад, «he was harassed online» можна перекласти як «його переслідували онлайн» або «він зазнав переслідувань в інтернеті». У

першому варіанті акцент зміщується на дію, у другому – на її наслідки, що може вплинути на спосіб, в який модель інтерпретує кіберзалякування та визначає його тип.

Зміна тональності тексту також є важливим фактором, наприклад, в англійській мові фраза «stop being so sensitive» звучить нейтрально або навіть вживається в дружньому контексті, тоді як український варіант «припини бути таким чутливим» виглядає різкіше, що може змінити класифікацію з нейтрального або жартівливого висловлювання на потенційне кіберзалякування.

Також існує ризик упередженості в перекладі, коли певні культурні чи соціальні особливості неправильно відтворюються. Наприклад, сленгове слово «karen» в англомовному середовищі є стереотипним образом людини, яка скаржиться на все підряд, але в українському контексті термін не має аналогів. Якщо перекласти його як «істерична жінка» або «скандальна пані», це спотворює вихідне значення і може вплинути на роботу моделі. Інший приклад – англійське «race card», що означає звинувачення в расизмі з метою отримання переваг. В українській мові немає прямого еквівалента і спроба дослівного перекладу викликає нерозуміння, що призодить до втрати частини інформації або неправильної класифікації.

Ще одним обмеженням є довжина текстових даних. Текст для аналізу має обмеження від 100 до 300 символів, що пов’язано з довжиною текстових зразків у датасетах, на яких навчались моделі машинного та глибокого навчання. Дуже короткі повідомлення розміром до 100 символів містять недостатньо інформації для коректного виявлення кіберзалякування, тексти понад 300 символів приводять до втрати контексту через обмеження моделі, що була навчена на текстових даних розміром 100–300 символів.

Для методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості діють окремі обмеження. В рамках дослідження методу визначено такий набір етичних аспектів, за якими виконується аналіз:

– віковий (вікові групи 0–19, 20–29, 30–39, 40–49, 50–100 років);

– цисгендерний (чоловіча стать, жіноча стать);

– релігійний (єврей, мусульманин, християнин, буддист).

Таке обмеження пов’язане з наявними датасетами, проте перелік може бути змінений або доповнений за потреби та наявності відповідних датасетів для моделей, які виконуватимуть розмітку датасетів для виявлення кіберзалякувань.

Для апробації методу оцінювання і коригування репрезентативності датасету за FATE-принципом справедливості в рамках дослідження використано два етичних аспекти – віковий та гендерний, на прикладі популяції України, проте коригування вхідного датасету може бути адаптоване під інші цільові пропорції популяції, наприклад, пропорції користувачів соціальних мереж та інші етичні аспекти.

Обмеженням методу виявлення кіберзалякувань у текстовому контенті є перелік типів кіберзалякувань, які здатна розпізнавати нейромережева модель. В рамках дослідження цей перелік складається з таких типів кіберзалякувань:

– вікове;

– гендерне;

– етнічне;

– релігійне;

– інші типи.

Такий вибір обумовлений складом навчального датасету, який використовувався для навчання моделі. Важливо зазначити, що перелік типів кіберзалякувань може змінюватися залежно від використовуваного датасету та потреб. Також на цей метод розповсюджуються і обмеження щодо мови та розміру текстового контенту для аналізу.

Для методу інтерпретації результатів виявлення кіберзалякувань діють ті самі обмеження, що описані вище для методу виявлення та класифікації типів кіберзалякувань у текстовому контенті.

Отже, методи виявлення та класифікації типів кіберзалякувань у текстовому контенті в межах дослідження обмежені мовою (українська), довжиною тексту, типами кіберзалякувань, що визначені навчальним датасетом.

2.3. Етичні та правові аспекти відповідального використання штучного інтелекту при виявленні кіберзалякувань у текстовому контенті

Відповідальне використання штучного інтелекту для виявлення кіберзалякувань у текстовому контенті вимагає ґрунтовного осмислення як етичних, так і правових аспектів, що виникають на перетині цифрових технологій, свободи вираження, права на приватність і необхідності захисту від кіберзалякувань в онлайн-середовищі. Етичні питання насамперед стосуються забезпечення справедливості, недискримінаційності та прозорості алгоритмічних рішень [50]. Алгоритми, що здійснюють виявлення кіберзалякувань у текстовому контенті, навчаються на обмежених або штучно сформованих датасетах, що часто містять приховану дискримінацію особистостей за статтю, віком, етнічністю чи релігією, або не враховують культурно-лінгвістичні особливості спілкування. В свою чергу це породжує ризик помилкової інтерпретації тональності, сарказму, сленгу або регіональних мовних відмінностей (наприклад, суржику в українській мові), що, в свою чергу, може приводити до хибної класифікації текстового контенту як такого, що не містить кіберзалякувань.

З етичного погляду важливо також гарантувати, що користувачі мають змогу оскаржити рішення, прийняті системами ШІ, і отримати пояснення щодо причин класифікації того чи іншого контенту як такого, що містить ознаки кіберзалякування. Прозорість алгоритмів – ключовий елемент довіри до цифрових систем, і її відсутність приводить до зниження довіри до результатів ШІ у соціальному дискурсі [74].

Правові аспекти регулювання ШІ у сфері протидії кіберзалякуванню пов’язані з дотриманням основоположних прав людини, передусім права на захист персональних даних і недоторканність приватного життя. Відповідно до Загального регламенту захисту даних Європейського Союзу, будь-яка система, що обробляє особисту інформацію, має діяти відповідно до принципів законності, справедливості. Це означає, що при створенні і впровадженні алгоритмів виявлення кіберзалякувань необхідно чітко визначати, які саме категорії даних обробляються, на якій підставі, з якою метою, а також забезпечити механізми контролю за обробкою таких даних та інформування суб’єктів даних про використання їх цифрового сліду в межах систем ШІ.

Особлива правова увага приділяється питанням відповідальності за наслідки рішень, прийнятих засобами ШІ, особливо в разі помилкової ідентифікації безневинного користувача як такого, що розповсюджує текстовий контент, що містить кіберзалякування, або у випадку ігнорування реального випадку кіберзалякування. Наразі законодавчі системи багатьох країн, включно з Україною, ще не мають чітко визначених норм щодо правового статусу систем прийняття рішень, що створює правову сферу у питаннях відповідальності за дії алгоритмів.

З етичної точки зору, помилкове визначення кіберзалякування системою штучного інтелекту порушує принципи справедливості та поваги до гідності особистості користувача. У разі хибнопозитивного результату, коли невинний користувач помилково ідентифікується як ініціатор прояву кіберзалякувань, відбувається безпідставне обмеження його свободи вираження поглядів, що суперечить принципу недискримінації. У випадку хибнонегативного результату, коли система не розпізнає наявне в текстовому контенті кіберзалякування, постраждала особа не отримує необхідного захисту, що суперечить етичному обов’язку гарантувати безпеку і добробут усіх учасників онлайн-комунікації. Саме тому етична модель впровадження таких систем передбачає обов’язкове

збереження за людиною права остаточного рішення, а також прозорість алгоритмічних процесів, наданих результатів, а також можливість їх оскарження.

З огляду на викладене, застосування штучного інтелекту для виявлення кіберзалякувань у текстовому контенті має базуватися на основах соціальної, етичної та правової відповідальності. Ефективне функціонування таких систем можливе лише за умови інтеграції етичних засад онлайн-комунікації, правових регламентів і критичного осмислення потенційних соціальних наслідків автоматизованого втручання.

2.4. Інформаційна модель кіберзалякування

У контексті задачі виявлення кіберзалякувань необхідно навести формальні позначення. Інформаційна модель кіберзалякування є інформаційним поданням його сутності, достатнім для розв’язання задачі виявлення та класифікації кіберзалякувань за запропонованим підходом.

Для задачі виявлення та класифікації кіберзалякувань, інформаційну модель кіберзалякування *CB* можна подати у вигляді:

$$CB = \{TT,\ OC,\ RC,\ OTC,\ TC',\ VA,\ MetaData\ \}, \tag{2.1}$$

де *TT* – тестовий текст для аналізу; *OC* – числова оцінка наявності кіберзалякувань в тексті *TT*; *RC* – рівень кіберзалякувань в тексті *TT*; *OTC* – множина числових оцінок сили прояву кожного з досліджуваних типів кіберзалякувань *TC* ($|TC| = |OTC|$); *TC′* – множина наявних у тексті типів кіберзалякувань ($TC' \subset TC$, де *TC* – множина всіх досліджуваних типів кіберзалякувань); *VA* – множина візуальних подань інтерпретації результатів виявлення кіберзалякувань; *MetaData* – множина метаданих моделі кіберзалякування *CB*.

У (2.1) метадані моделі кіберзалякувань *CB* можна подати у вигляді:

$$MetaData = \{MetaData1,\ MetaData2,\ MetaData3\}, \tag{2.2}$$

де *MetaData*1 – множина метаданих для оцінювання репрезентативності датасетів, *MetaData*2 – множина метаданих для виявлення кіберзалякувань; *MetaData*3 – множина метаданих для інтерпретації результатів.

Множина метаданих *MetaData*2 для виявлення та класифікації кіберзалякувань у текстовому контенті у (2.2) містить:

$$MetaData2 = \{TT_{2,3}, NN_{cbt}, NN_{cbd}\}, \quad (2.3)$$

де $TT_{2,3}$ – текст після препроцесингу $TT \rightarrow TT_{2,3}$ (ідентичний для *MetaData*2 та *MetaData*3); NN_{cbt} – нейромережева модель для виявлення кіберзалякувань (бінарна класифікація); NN_{cbd} – нейромережева модель для класифікації кіберзалякувань (мультилейблова класифікація).

NN_{cbd} у дослідженні виконує мультилейблову класифікацію типів кіберзалякувань, що формують множину $TC = CBT = \{$*«вікове», «гендерне», «релігійне», «етнічне», «інші типи»*$\}$.

Множина метаданих *MetaData*3 для інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами у (2.2) подається у вигляді:

$$MetaData3 = \{TT_{2,3}, W, IM_{Interpet}\}, \quad (2.4)$$

де $IM_{Interpet}$ – інтерпретаційна модель для пояснення результатів класифікації NN_{cbd} (візуальна аналітика), W – множина значущих слів для типів кіберзалякувань у тексті $TT_{2,3}$.

У результаті застосування інтерпретаційної моделі $IM_{Interpet}$ формуються кортежі вигляду $W=\{(w_1,\beta_1),(w_2,\beta_2),...,(w_n,\beta_n)\}$, де w_i – i-те слово; β_i – вага i-го слова; n – кількість значущих слів у тексті $TT_{2,3}$ для обраного типу кіберзалякувань. Кортежі наведеного вигляду використовується для створення множини графічних подань інтерпретації результатів виявлення кіберзалякувань *VA*.

Множина метаданих *MetaData*1 для оцінювання репрезентативності датасетів у (2.2) містить складові, призначені для вирішення проблеми одержання репрезентативного, недискримінаційного за FATE-принципом справедливості

текстового датасету можна подати у рамках інформаційної моделі такого вигляду:

$$Metadata1 = \{DS, DS', C, A, M\}, \quad (2.5)$$

де DS – текстовий датасет для аналізу та коригування; DS' – текстовий датасет після коригування; C – множина класів предметної області датасету; A – множина етичних аспектів FATE-принципу справедливості; M – множина навчених моделей машинного навчання (окрема для кожного етичного аспекту).

У (2.5) початковий датасет DS та відкоригований датасет DS' можуть бути подані у вигляді:

$$DS = \{D \cup MDS\}, \quad (2.6)$$

$$DS' = \{D' \cup MDS'\}, \quad (2.7)$$

де D – множина елементів датасету DS; MDS – множина метаданих датасету DS; D' – множина елементів датасету DS'; MDS' – множина метаданих датасету DS'.

Кожен елемент множини елементів датасету D у (2.6) та кожен елемент множини елементів датасету D' у (2.7) є кортежем такого вигляду:

$$d = d' = (text, c_x, AC_x), \quad (2.8)$$

де атрибут *text* – текстовий вміст елементу d або d'; c_x – клас предметної області датасету, до якого належить елемент, $c_x \in C$; AC_x – множина класів приналежності елементу датасету до етичних аспектів.

Таким чином, в (2.8) c_x та AC_x є маркуванням (розміткою) контенту елемента *text*.

У (2.8) множина класів приналежності елементу датасету DS або DS' у (2.5) до етичних аспектів A_x подається у вигляді кортежа:

$$A_x = (a_{1,x}, a_{2,x}, \ldots, a_{k,x}), \quad (2.9)$$

де a_x – класи приналежності елементу до етичних аспектів; k – кількість етичних аспектів, що підлягають аналізу, $k = |A_x|$.

При цьому в (2.9) відповідно до (2.5), $A_x \subset A$, й класи приналежності елементів датасету до етичних аспектів є елементами відповідних множин, унікальних для кожного з етичних аспектів:

$$a_{1,x} \in A_1,\ a_{2,x} \in A_2,\ \ldots,\ a_{k,x} \in A_k, \tag{2.10}$$

$$A_1 \cup A_2 \cup \ldots \cup A_k = A \tag{2.11}$$

До множини метаданих *MDS* датасету *DS* у (2.6) належать:

$$MDS = \{n_{DS},\ AN_{DS},\ AT_{DS},\ n'_{DS},\ AN'_{DS},\ AT'_{DS}\}, \tag{2.12}$$

де n_{DS} – кількість елементів у D, $n_{DS} = |D|$; AN_{DS} – множина кількостей елементів датасету, що належать кожному класу кожного етичного аспекту з A_x; AT_{DS} – множина наявних пропорцій елементів для кожного класу відносно загальної кількості за кожним етичним аспектом з A_x; n'_{DS} – цільова кількість елементів у D'; AN'_{DS} – множина цільових кількостей елементів датасету, що належать кожному класу кожного етичного аспекту з A_x; AT'_{DS} – множина цільових пропорцій елементів для кожного класу відносно загальної кількості за кожним етичним аспектом з A_x.

При цьому, у (2.12) кожен елемент $an_{DS,i}$ множини AN_{DS} відповідає окремому i-му етичному аспекту й подається кортежом такого вигляду:

$$an_{DS,i} = (n_{DS,i,1},\ n_{DS,i,2},\ \ldots,\ n_{DS,i,j},\ \ldots,\ n_{DS,i,k}), \tag{2.13}$$

де $n_{DS,i,1}$ – кількість елементів у датасеті 1-го класу i-го етичного аспекту; $n_{DS,i,2}$ – кількість елементів у датасеті 2-го класу i-го етичного аспекту; $n_{DS,i,j}$ – кількість елементів у датасеті j-го класу i-го етичного аспекту; k – кількість класів i-го етичного аспекту.

Аналогічно до (2.13), у (2.12) пропорції елементів $at_{DS,i}$ i-го етичного аспекту подаються кортежом такого вигляду:

$$at_{DS,i} = (t_{DS,i,1},\ t_{DS,i,2},\ \ldots,\ t_{DS,i,j},\ \ldots,\ t_{DS,i,k}), \tag{2.14}$$

де $t_{DS,i,1}$ – відношення кількості елементів у датасеті 1-го класу i-го етичного аспекту до загальної кількості елементів у датасеті; $t_{DS,i,2}$ – відношення кількості елементів у датасеті 2-го класу i-го етичного аспекту до загальної кількості

елементів у датасеті; $t_{DS,i,j}$ – відношення кількості елементів у датасеті i-го класу i-го етичного аспекту до загальної кількості елементів у датасеті.

При цьому, для значень (2.13) та (2.14) відповідно до (2.12) для кожного i-го етичного аспекту справджується рівність:

$$n_{DS,i,1} + n_{DS,i,2} + \ldots + n_{DS,i,k} = n_{DS}, \quad (2.15)$$

$$t_{DS,i,1} + t_{DS,i,2} + \ldots + t_{DS,i,k} = 1. \quad (2.16)$$

На відміну від (2.12), до множини метаданих *Metadata′* датасету *DS′* у (2.7) належать:

$$MDS' = \{n''_{DS}, AN''_{DS}, AT''_{DS}\}, \quad (2.17)$$

де n''_{DS} – фактично одержана в результаті коригування кількість елементів у D', $n''_{DS} = |D|$; AN''_{DS} – множина фактично одержаних в результаті коригування кількостей елементів датасету, що належать кожному класу кожного етичного аспекту з A_x; AT''_{DS} – множина фактично одержаних в результаті коригування пропорцій елементів для кожного класу відносно загальної кількості за кожним етичним аспектом з A_x.

При цьому, у (2.12) та (2.17) для AN'_{DS} та AN''_{DS} справджуються (2.3) та (2.15), а для AT'_{DS} та AT''_{DS} справджуються (2.14) та (2.16).

Таким чином, відповідно до (2.8), (2.10) та (2.11), текстовий датасет D має кількість елементів $n = n_{DS} = |D|$ та може бути поданий у вигляді:

$$D = \{d_1, d_2, \ldots, d_n,\}, d_i = (text_i, c_i, A_1, A_2, \ldots, A_m), i = \overline{1, \ldots, n}, \quad (2.18)$$

де $C = \{c_1, c_2, \ldots, c_k\}$, тут k – кількість класів датасету D; m – кількість етичних аспектів.

Відповідно до (2.10) – (2.14), розв’язання задачі спрямоване на одержання датасету D', який містить загальну кількість елементів $n' = n'_{DS} = |D'|$, кількісно збалансованих за етичними аспектами A_i з множини етичних аспектів A:

$$A = \{A_1, A_2, \ldots, A_m\}, A_i = \left(C_i, T_{ij}\right), i = \overline{1, \ldots, m}, \quad (2.19)$$

де кожен аспект A_i містить класи C_i та цільові пропорції класів T_{ij} для кожного елементу класу C; C – множина класів етичного аспекту A_i , $C = \{c_1, c_2, \ldots, c_j\}$; j – кількість класів етичного аспекту A_i.

Для балансування датасету за кожним етичним аспектом необхідно використати навчені або навчити відповідну кількість моделей класифікаторів, якими можуть бути як моделі глибокого навчання, наприклад, BERT [105], LSTM [110], GRU [111], так і моделі машинного навчання Logistic Regression [112], Naive Bayes [115], Support Vector Machines (SVM) [117], k-Nearest Neighbors (k-NN) [119] тощо, й відповідно до (2.5) множина навчених моделей класифікаторів M подається у вигляді:

$$M = \{M_1, M_2, \ldots, M_m\}, m = |D|. \quad (2.20)$$

У рамках запропонованої інформаційної моделі, необхідно виконати перетворення $D \Rightarrow D'$ з умовою максимальної відповідності $n''_{DS} \rightarrow n'_{DS}$, $AN''_{DS} \rightarrow AN'_{DS}$ та $AT''_{DS} \rightarrow AT'_{DS}$.

Розроблена інформаційна модель кіберзалякування забезпечує подання його сутності в інформаційному обсязі, достатньому для розв’язання задачі виявлення та класифікації кіберзалякувань за запропонованим підходом, що містить реалізацію розроблених методів оцінювання і коригування репрезентативності датасетів, виявлення кіберзалякувань та інтерпретації результатів.

2.5. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості дозволяє вирішити протиріччя між можливістю точного виявлення кіберзалякувань у текстовому контенті та відсутністю довіри до навчальних даних, які не можуть гарантувати репрезентативність результатів через відсутність перевірки та можливостей приведення навчальних даних до

репрезентативного вигляду. Як наслідок застосування запропонованого методу, забезпечується недискримінація за віковою, гендерною та релігійною приналежністю, що дозволяє підвищити якість навчання класифікаторів для виявлення кіберзалякувань.

2.5.1. Схема методу

Концепція оцінювання та коригування репрезентативності датасету. У дослідженні пропонується звести задачу побудови репрезентативного за етичними аспектами FATE-принципу справедливості датасету до задачі багатокритеріальної оптимізації. Задача оптимізації полягає у мінімізації відхилення між поточними та бажаними співвідношеннями класів, враховуючи обмеження на кількість зразків у класах і можливостей генерації синтетичних даних.

Вхідні дані: текстовий датасет *DS*, множина етичних аспектів *A*, вимоги до репрезентативного розподілу *DS′*.

Мета задачі: створення репрезентативної вибірки за всіма етичними аспектами, яка досягає цільових пропорцій класів для кожного етичного аспекту $D \Rightarrow D'$.

Змінні: x_{ij} – кількість зразків класу C_j в аспекті A_i після видалення та аугментації.

Обмеження задачі:

1) сума всіх зразків класів в межах одного аспекту дорівнює цільовій кількості зразків для цього аспекту (2.21):

$$\sum_{j=1}^{n_i} x_{ij} = n', \forall i \in \{1,2,\dots,m\}, \tag{2.21}$$

де n_i – кількість класів в аспекті A_i;

2) кількість зразків для кожного класу повинна відповідати цільовій пропорції класів:

$$\frac{x_{ij}}{n'} \approx T_{ij}, \forall i \in \{1,2,\dots,m\}, \forall j \in \{1,2,\dots,n_i\}; \quad (2.22)$$

3) розрахункова кількість зразків не може бути від'ємною:

$$x_{ij} \geq 0, \forall i \in \{1,2,\dots,m\}, \forall j \in \{1,2,\dots,n_i\}; \quad (2.23)$$

4) можливість додавання нових зразків повинна відповідати можливостям генерації нових даних для кожного класу та аспекту:

$$x_{ij} \leq G_{ij}, \forall i \in \{1,2,\dots,m\}, \forall j \in \{1,2,\dots,n_i\}, \quad (2.24)$$

де G_{ij} – максимально можлива кількість зразків класу C_j в аспекті A_i, яку можна додати.

Цільовою функцією є мінімізація відхилення між поточними та бажаними співвідношеннями для всіх аспектів одночасно з урахуванням обмежень (2.21) – (2.24):

$$\sum_{i-1}^{m}\sum_{j-1}^{n_i}\left|\frac{x_{ij}}{N_{target}} - T_{ij}\right| \to min. \quad (2.25)$$

Далі наведено кроки методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості відповідно до поставленої оптимізаційної задачі формування репрезентативного датасету (2.25).

Схема та кроки методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості. Метод передбачає не тільки аналіз на репрезентативність, а й коригування репрезентативності датасету для подальшого виявлення кіберзалякувань у текстовому контенті. Слід зауважити, що безпосереднє доповнення датасету зразками, згенерованими, наприклад, за методикою Synthetic Minority Over-sampling Technique (SMOTE) [120], не є оптимальним, оскільки таке репрезентативне коригування датасету за одним аспектом призведе до

нерепрезентативного подання даних за іншими етичними аспектами. Згідно з (2.25), необхідне багатокритеріальне коригування датасету шляхом видалення зразків та аугментації за кількома етичними аспектами одночасно. Наприклад, необхідно сформувати текстовий датасет, репрезентативний за двома етичними аспектами – гендерним і віковим. Виявивши нерепрезентативне подання за віковим аспектом, необхідно доповнити датасет таким чином, щоб не порушити репрезентативність датасету за гендерним аспектом, таким чином необхідно розв'язати задачу (2.25) оптимізації формування репрезентативного датасету за усіма обраними етичними аспектами одночасно.

Варто зауважити, що в межах цієї роботи при аналізі гендерного етичного аспекту розглядається цисгендерна група. Схему методу наведено на рис. 2.4.

Рис. 2.4 – Кроки методу аналізу та формування репрезентативних вибірок текстових даних

Вхідними даними методу є датасет DS для аналізу, що за (2.6) та (2.12) містить цільову кількість елементів n'_{DS}, множину етичних аспектів A з підмножинами класів, цільові пропорції AT_{DS} класів та кількості AN'_{DS} елементів у класах етичних аспектів, відповідно навчена множина моделей M для кожного етичного аспекту з A, яка для навчання використовує збалансовані вибірки для кожного етичного аспекту.

На кроці 1 здійснюється попередня обробка вибірки текстових даних у $D \subset DS$, а саме видалення неінформативних фрагментів тексту, таких як знаки пунктуації, цифри та спеціальні символи [116]. Видалення смайлів не виконується, оскільки в багатьох випадках їхнє включення в аналіз дозволяє покращувати точність моделей машинного навчання, що використовуються для класифікації текстів за емоційним або настроєвим змістом [121].

На кроці 2 здійснюється аналіз репрезентативності вибірки текстових даних з урахуванням етичних аспектів. Спершу необхідно здійснити векторизацію й класифікацію кожного елемента $\forall d \in D$ вибірки даних, використовуючи окремі моделі машинного навчання $m \in M$ за кожним з етичних аспектів $A_i \in A$. Визначаються наявні пропорції класів AN_{DS} та AT_{DS} для кожного з етичних аспектів. Обраховуються кількості нестачі або надлишку елементів кожного класу за кожним з етичних аспектів. Після цього виконується аналіз достатності даних у вибірці для аугментації (мінімальна наявність зразків відповідних класів тощо).

Крок 3 передбачає репрезентативне коригування вибірки даних з урахуванням етичних аспектів. Коригуванням є видалення та аугментація зразків.

Видалення надлишкових елементів кожного класу за кожним з етичних аспектів відбувається з мінімальною шкодою для інших розподілів, для чого розв’язується оптимізаційна задача вибору надлишкових елементів в рамках (2.25), які мають бути видалені для досягнення цільових пропорцій класів.

Операція аугментації виконується для створення нових елементів за допомогою одного з відомих способів [116]. Створюються вимоги у вигляді

потрібної комбінації класів кожного з етичних аспектів для кожного нового елементу, для чого розв’язується оптимізаційна задача формування вимог до відсутніх елементів в рамках (2.25).

Вихідними даними методу є текстовий датасет $D' \subset DS'$, що має необхідний обсяг n'_{DS} та збалансований згідно з необхідними пропорціями AT'_{DS} за обраними етичними аспектами $A_x \subset A$.

Кінцевою метою коригування датасетів є забезпечення балансу не тільки за FATE-принципом справедливості (етичні аспекти), а й за кількістю зразків у класах (види кіберзалякувань), що виконується аналогічним чином.

Виконання кроків запропонованого методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості дозволяє формувати датасети, які є репрезентативними і недискримінаційними та відображають пропорційне до реальних демографічних підгруп популяції представлення зразків датасету, що впливатиме на точність та прозорість навчання моделей машинного навчання для розв’язання різноманітних задач.

Отримання типових моделей машинного навчання для етичних аспектів. Для формування множини навчених моделей машинного навчання *M*, які є окремими для кожного етичного аспекту з множини аспектів *A*, необхідно навчити кожну модель класифікатора, що аналізуватиме репрезентативність вхідної текстового датасету *DS* згідно з кроком 2 на рис. 2.4. Для отримання таких класифікаторів, що і будуть формувати множину навчених моделей машинного та глибокого навчання, необхідно виконати кроки, що подані на рис. 2.5.

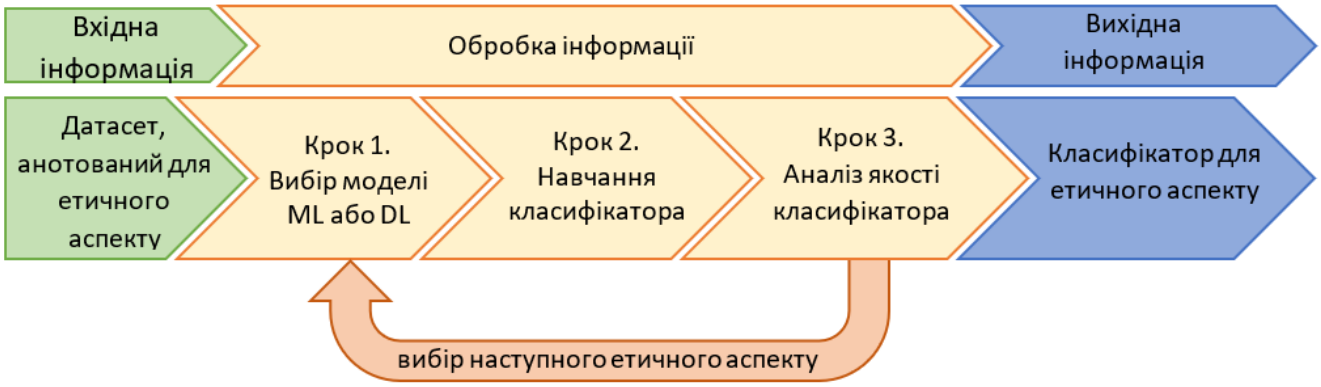


Рис. 2.5 – Процес отримання типових моделей машинного навчання для етичних аспектів

Для отримання моделей машинного навчання з множини етичних аспектів вхідними даними є датасет, анотований для окремого етичного аспекту, який розглядається. Першим кроком є вибір моделі для навчання класифікатора. Для таких цілей використовуються як моделі глибокого навчання, наприклад, BERT, GPT, LSTM, GRU, так і класифікатори Logistic Regression, Naive Bayes, SVM, k-NN тощо. Після цього відбувається навчання класифікатора за вибраною моделлю машинного або глибокого навчання на анотованому датасеті для етичного аспекту.

Однією з важливих особливостей глибокого навчання є його здатність до багаторівневої абстракції даних, де кожен шар нейронної мережі перетворює вхідні дані в набір більш абстрактних характеристик, що дозволяє моделям глибокого навчання вирішувати складні завдання, включаючи аналіз тексту [128].

Отже, наведено послідовність кроків отримання типових моделей машинного навчання для етичних аспектів. Варто зазначити, що наведені кроки можуть застосовуватись і для отримання моделей виявлення та класифікації типів кіберзалякувань у текстовому контенті.

2.5.2. Демонстрація роботи методу

Для демонстрації роботи методу наведено приклад з використанням датасету із 10 000 анонімізованих коментарів із соціальної мережі. Етичними аспектами, що враховувалися в процесі формування датасету, були гендер (чоловік / жінка) та вік (молодь до 25 років і дорослі 26–60 років). Цільовими пропорцями для цих категорій встановлено 50 % для кожної статі та 30 % – молоді проти 70 % дорослих.

На першому кроці відбулася попередня обробка вхідного датасету для виявлення кіберзалякувань. Було вилучено всі неінформативні елементи, зокрема фрагменти, що складалися виключно із знаків пунктуації, смайлів, цифрових кодів або HTML-міток, текстові зразки розміром до 100 символів та понад 300. Також видалено порожні записи або повідомлення, що не містили тексту, що несе смислове навантаження. Приклади вилучених елементів: «*))))))))))*», «*!!!*», «*123456*», «* *». У результаті цієї обробки з початкових 10 000 залишилося 9 800 валідних текстів для подальшого аналізу.

Таблиця 2.1

Оцінка етичної репрезентативності вибірки за окремими етичним аспектами

Категорія	**Підгрупа**	**Кількість**	**Фактичний %**	**Цільовий %**	**Відхилення, %**	**Тип відхилення**
Стать	Чоловіки	6860	70	50	20	Надлишок
	Жінки	2940	30	50	–20	Нестача
Вік	Молодь (≤25 років)	4900	50	30	20	Надлишок
	Дорослі (26–60 років)	4900	50	70	–20	Нестача

На другому кроці здійснено аналіз репрезентативності наявного датасету з урахуванням зазначених етичних аспектів. Тексти були попередньо оброблені та за допомогою моделей машинного навчання кожен елемент класифіковано за гендером і віковою категорією відповідно до вказаних етичних аспектів. У результаті аналізу виявлено суттєве порушення пропорцій: чоловіки становили 70 % від загальної кількості, жінки – лише 30 %, тоді як молоді та дорослі представлені порівну, що не відповідало цільовому розподілу (табл. 2.1).

Аналіз показав, що найбільш недостатньо представленою групою є жінки віком від 26 до 60 років, тоді як чоловіки до 25 років, навпаки, були надмірно представлені (табл. 2.2). Такий дисбаланс призводитиме до дискримінаційного навчання моделі для завдання виявлення кіберзалякувань.

Таблиця 2.2

Аналіз нестачі / надлишку за підгрупами за обома етичними аспектами одночасно

Група	**Кількість**	**Цільова пропорція, %**	**Відхилення, %**
Чоловіки до 25 років	3400	15	25
Жінки до 25 років	1500	15	0
Чоловіки 26–60 років	3460	35	–5
Жінки 26–60 років	1440	35	**–20**

Третій крок передбачає репрезентативне коригування вибірки. Для цього було розв’язано оптимізаційну задачі видалення надлишкових елементів та аугментації малопредставлених груп. З вибірки було вилучено 900 текстів, що належали до надлишкової групи чоловіків до 25 років. Водночас здійснено штучне

розширення вибірки для групи жінок віком 26–60 років шляхом SMOTE-балансування. Усього для забезпечення збалансованості додано 720 нових елементів для найбільш малопредставленої групи та ще 180 елементів для додаткового вирівнювання інших груп.

У результаті датасет містив 9 000 текстів, збалансованих згідно з етичними аспектами та цільовими пропорціями: однакова кількість чоловіків і жінок, а також співвідношення молоді та дорослих відповідно до цільових пропорцій – 30 % і 70 % (табл. 2.3).

Таблиця 2.3

Розподіл згідно з цільовими пропорціями

Група	**Кількість**
Чоловіки до 25 років	1350
Жінки до 25 років	1350
Чоловіки 26–60 років	3150
Жінки 26–60 років	3150

Таким чином, було досягнуто цільових пропорцій та відповідно до цього недискримінаційного представлення соціальних груп у вхідному датасеті для оцінки та коригування репрезентативності. Датасет є репрезентативним і недискримінаційним за FATE-принципом справедливості за такими етичними аспектами, як вік та гендер, тому може використовуватись для навчання моделей машинного навчання для виявлення кіберзалякувань у текстововму контенті.

2.6. Метод виявлення кіберзалякувань у текстовому контенті

2.6.1. Схема методу

Метод виявлення кіберзалякувань у текстовому контенті дозволяє вирішити проблему низької точності їх класифікації внаслідок того, що деякі типи кіберзалякувань мають спільні ознаки, й для підвищення точності класифікації кіберзалякувань вимагається навчання нейромережевої моделі виключно даними з високим рівнем, тоді як для визначення рівня кіберзалякувань у тексті має використовуватись клас даних без кіберзалякувань [106]. Розроблений метод відрізняється від існуючих двоетапним виявленням кіберзалякувань, що полягає у нейромережевій ідентифікації наявності кіберзалякувань та подальшій нейромережевій мультилейбловій класифікації їх окремих типів. Застосування методу виявлення кіберзалякувань у текстовому контенті дає можливість підвищити точність та якість виявлення кіберзалякувань [107].

Метод виявлення та класифікації кіберзалякувань призначений для аналізу текстового контенту з метою виявлення та класифікації типів кіберзалякувань. Схему методу наведено на рис. 2.6. Схема демонструє процес виявлення та класифікації кіберзалякувань у текстовому контенті, що має три кроки.

Перший крок – це попередня обробка тексту для аналізу, де $TT \rightarrow TT_{2,3}$. Вхідними даними на цьому етапі є тестовий текст TT для виявлення кіберзалякувань. Вхідний тестовий текст очищується від зайвих символів, таких як знаки пунктуації та зайві пробіли. У результаті обробки отримується текст у чистому вигляді $TT_{2,3}$, який перетворюється на векторне представлення для подальшої роботи з нейромережею BiLSTM.

На другому кроці здійснюється аналіз тексту з метою виявлення кіберзалякувань за допомогою навченої моделі NN_{cbt} BiLSTM. Нейромережева модель NN_{cbt} навчена на репрезентативному датасеті DS' [108].

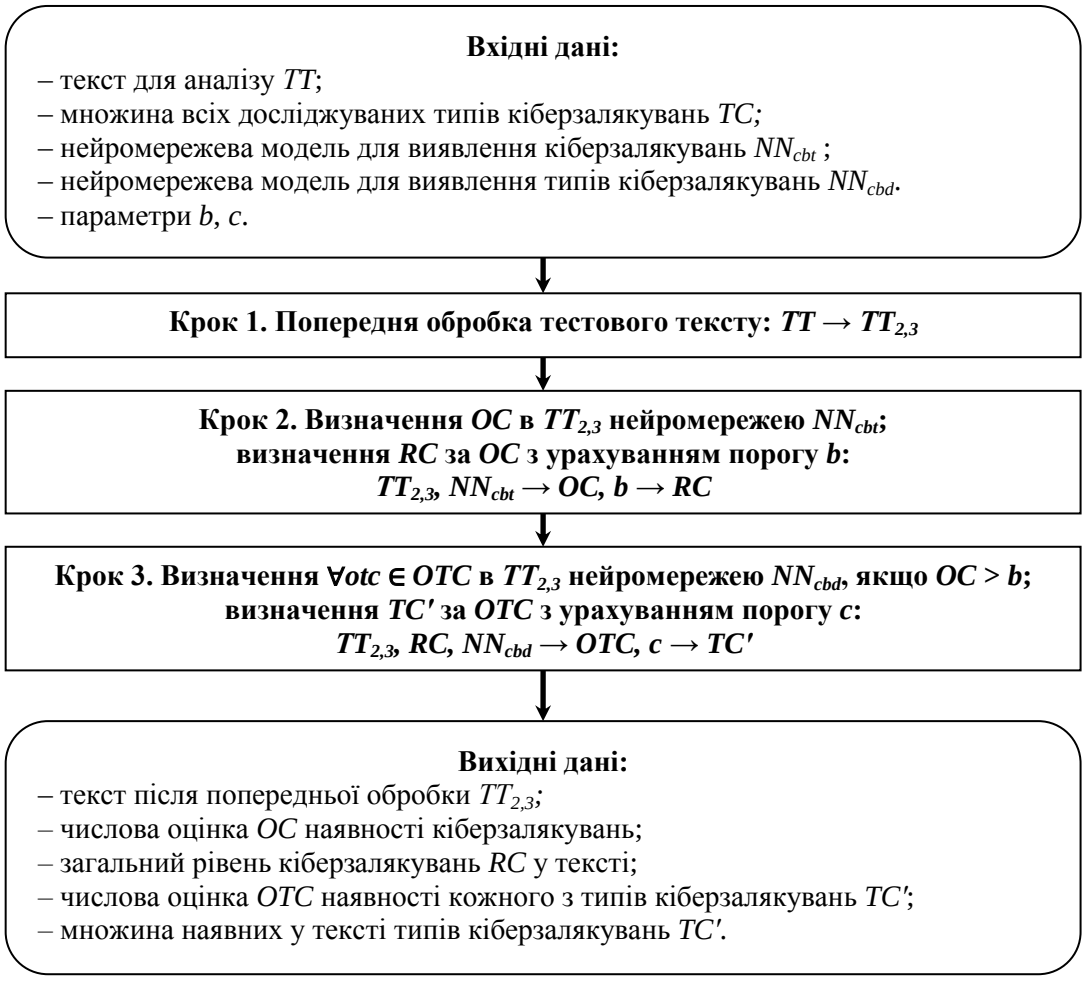


Рис. 2.6 – Схема методу виявлення і класифікації кіберзалякувань у текстовому контенті

Вибір BiLSTM обґрунтовується її перевагами над стандартною RNN та простою LSTM при обробці великих обсягів навчальних даних. Завдяки двонаправленій структурі ця модель аналізує контекст слова як у попередньому, так і в зворотному напрямках, що є особливо важливим для розпізнавання натяків або завуальованих форм кіберзалякувань. Ця модель оцінює текст $TT_{2,3}$ на предмет наявності ознак кіберзалякування та обраховує числову оцінку наявності кіберзалякувань *OC* в тексті $TT_{2,3}$.

Якщо числова оцінка *OC* перевищує поріг *b,* що має рекомендоване значення 0,5, то текст класифікується як такий, що містить ознаки кіберзалякувань, і виконується наступний крок мультилейблової класифікації типів кіберзалякувань з множини *TC* нейромережевою моделлю NN_{cbd}. Значення b = 0,5 встановлено емпіричним шляхом при проведенні дослідження ефективності запропонованого методу (п. 4.2).

Третій крок передбачає визначення типів кіберзалякувань *TC'* з множини *TC* у тексті $TT_{2,3}$. Для цього використовується модель NN_{cbd} BERT, яка навчена для виявлення типів кіберзалякувань з множини *TC* і визначає числову оцінку *OTC* наявності кожного з досліджуваних типів кіберзалякувань *TC* у тексті $TT_{2,3}$, що аналізується. Використання BERT для цієї задачі зумовлене її вже реалізованим попереднім навчанням на великих корпусах текстів, що дозволяє їй ефективно аналізувати контекст кожного слова. Додатково модель була дотренована на типи кіберзалякувань з множини *TC*. Це особливо важливо для точної класифікації кіберзалякувань, де значення висловлювань може змінюватися залежно від контексту [98].

Завдяки трансформерній архітектурі BERT демонструє високу здатність до узагальнення навіть у випадках нерівномірного розподілу даних між класами. Крім того, вона підтримує концепцію «zero-shot learning», що дозволяє прогнозувати нові або недостатньо представлені категорії на основі знань, отриманих у процесі попереднього навчання [99, 102]. На цьому кроці здійснюється мультилейблова класифікація типів кіберзалякувань *TC*, адже віднесення текстового зразка лише до одного типу кіберзалякування, як це відбувається при багатокласовій класифікації, є недостатнім. Оскільки один текстовий зразок може містити відразу кілька типів кіберзалякувань з множини *TC*, то доцільним є використання мультилейблової класифікації, яка дозволяє більш точно ідентифікувати всі типи кіберзалякувань *TC*.

На виході методу надається токенізований текст $TT_{2,3}$, числова оцінка *ОС* наявності кіберзалякувань у тексті *ТТ*, рівень кіберзалякувань *RC* у тексті *ТТ*, числова оцінка *ОТС* наявності кожного з досліджуваних типів кіберзалякувань *ТС* та множина *ТС′* наявних у тексті типів кіберзалякувань. Загальна оцінка *ОС* та оцінки *ОТС* для наявних типів кіберзалякувань *ТС′* у тексті *ТТ* надаються бінарною та мультилейбловою моделями у вигляді числа в проміжку [0, 1].

Наведений метод виявлення та класифікації кіберзалякувань дозволяє підвищити точність та якість виявлення кіберзалякувань, що досліджено у п. 4.2. Підвищення точності полягає у застосуванні двоетапного виявлення кіберзалякувань, що полягає у нейромережевій ідентифікації наявності кіберзалякувань і подальшій нейромережевій мультилейбловій класифікації окремих типів кіберзалякувань [109]. Підвищення якості виявлення кіберзалякувань полягає у навчанні класифікаторів для виявлення кіберзалякування з використанням репрезентативного датасету *DS′,* що отриманий в результаті застосування методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості.

2.6.2. Демонстрація роботи методу

Для демонстрації роботи методу наведено приклад з використанням текстового зразка «*Чого ти ще сюди лізеш, стара карга*? *Ніхто не питає думки баби, якій давно пора на пенсію*! *Йди краще пиріжки пекти*!». Разом з текстом на вхід також подаються дві нейромережеві моделі: одна для визначення наявності кіберзалякування, а інша – для класифікації його типів. Обидві моделі навчені з використанням репрезентативних датасетів, що отримані шляхом застосування методу оцінювання і коригування репрезентативності датасету за FATE-принципом справедливості. Окрім того, вказуються порогові значення, які визначають відповідно до отриманої числової оцінки наявності кіберзалякувань,

загальний рівень кіберзалякувань в тексті (присутній чи ні). Нехай таким порогом буде значення 0,5 для наявності кіберзалякування та 0,2 – для наявності кожного з типів кіберзалякування.

Перший крок – попередня обробка тексту. З нього прибираються розділові знаки, зайві пробіли, приводиться все до нижнього регістру, лематизуються, після цього текст перетворюється на формат, зручний для подальшого аналізу. З вхідного тексту отримано «*чого ти ще сюди лізеш стара карга ніхто не питає думки баби якій давно пора на пенсію йди краще пиріжки пекти*».

На другому кроці оброблений текст передається першій нейромережевій моделі, яка визначає, чи є в цьому тексті кіберзалякування. Модель оцінює його рівень та видає числову оцінку наявності кіберзалякувань у текстовому контенті. Далі визначається загальний рівень кіберзалякувань в тексті – якщо оцінка їх наявності вища за встановлений поріг (наприклад, 50 %), то робиться висновок про те, що кіберзалякування в тексті є. Модель визначає числову оцінку наявності кіберзалякувань 89,2 %, що більше, ніж встановлений поріг, а, отже, встановлено, що текст містить кіберзалякування.

Після цього відбувається наступний крок аналізу – класифікація типів кіберзалякування. Для цього використовується інша нейромережа, яка аналізує, які типи кіберзалякувань присутні у текстовому контенті. Модель надає оцінки для кожного типу:

– вікове – 91 %;
– гендерне – 87 %;
– етнічне – 18 %;
– релігійне – 4 %;
– інший тип – 16 %.

Найбільші оцінки отримали такі типи кіберзалякувань, як вікове і гендерне, їх значення перевищують поріг 20 %, тому робиться висновок про те, що типи наявні в тексті.

У результаті вихідними даними є: текст після попередньої обробки, який використовується для інтерпретації результатів кіберзалякувань, числова оцінка ймовірності наявності кіберзалякування 89,2 %, загальний рівень – «присутнє», числові значення ймовірностей для кожного типу кіберзалякування: вікове – 91 %, гендерне – 87 %, етнічне – 18 %, релігійне – 4 %, інший тип – 16 %; висновок про те, що в текстовому зразку, згідно із встановленим пороговим значенням 20% присутні такі типи кіберзалякувань, як вікове та гендерне.

2.7. Метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті

2.7.1. Схема методу

Метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті дозволяє вирішити проблему зниження довіри до результатів нейромережевих рішень з виявлення кіберзалякувань внаслідок їх низької пояснненності. Запропонований метод інтерпретації результатів виявлення кіберзалякувань відрізняється від існуючих можливістю надавати візуальні пояснення для мультилейблової класифікації виявлених типів кіберзалякувань в альтернативних поданнях [113, 114].

Схему методу інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту подано на рис. 2.7. Метод передбачає створення множини графічних подань інтерпретації результатів виявлення кіберзалякувань *VI* з використанням $IM_{Interpet}$.

Вхідними даними наведеного на рис. 2.7 методу є навчена модель NN_{cbd} для мультилейблової класифікації, яка здатна розпізнавати різні типи кіберзалякування з множини *CBT* [6]. Використовується інтерпретаційна модель $IM_{Interpet}$, яка дозволяє пояснити вплив окремих слів чи фраз на результат класифікації [129]. Перелік класів кіберзалякувань включає типи кіберзалякувань з

множини *CBT*, за якими модель здійснює класифікацію та виконується інтерпретація отриманих результатів. Вхідними даними також є токенізований текст $TT_{2,3}$, який класифіковано *RC* як такий, що містить кіберзалякування, тобто для якого числова оцінка наявності кіберзалякувань *OC* має значення вище порога *b* (рекомендується *b* = 0,5).

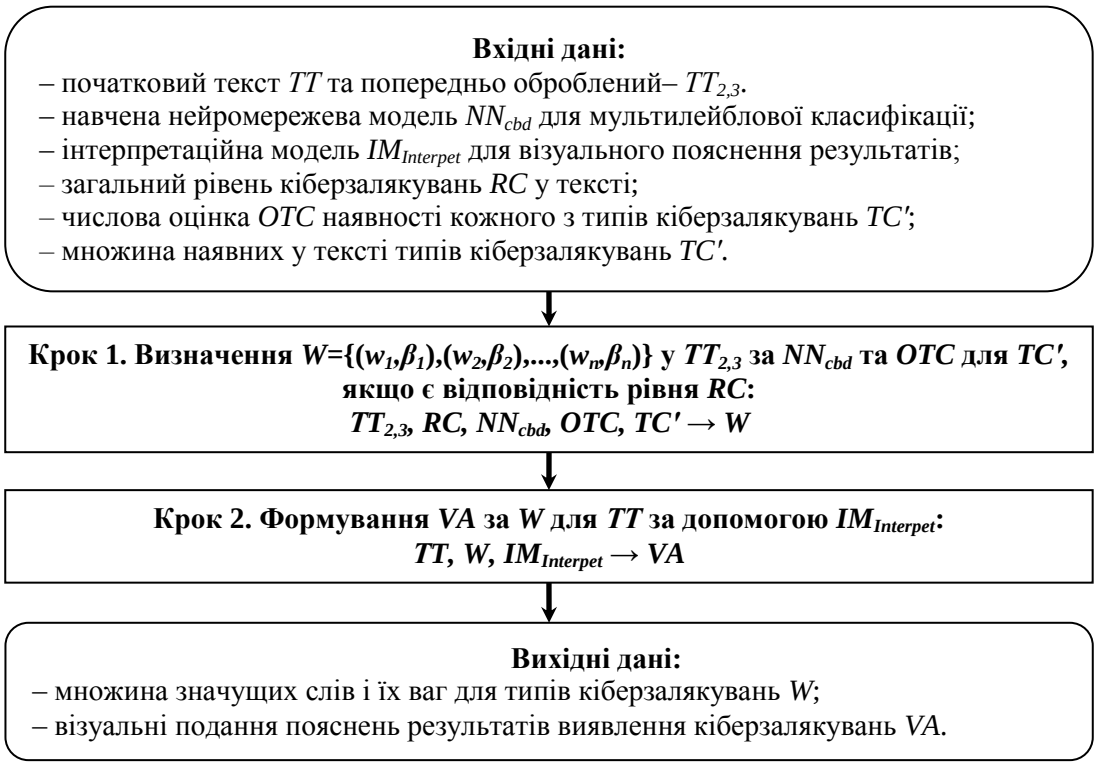


Рис. 2.7 – Схема методу інтерпретації результатів виявлення кіберзалякувань у текстовому контенті

На кроці 1 інтерпретаційна модель $IM_{Interpet}$, отримуючи токенізований тестовий текст $TT_{2,3}$ [130], обраховує показники важливості слів у контексті кожного типу кіберзалякування з множини *CBT*, таким чином формуючи кортеж *W*, в який входять слова w_i тестового тексту $TT_{2,3}$ та їх ваги β_i, що вказують, які із слів найбільше вплинули на класифікацію тексту за певним типом кіберзалякування [131, 8]. Варто зазначити, що вплив слів може бути як позитивним, так і негативним. Додатні ваги (позитивний вплив) вказують на

слова, що підсилюють ймовірність належності тексту до певного типу кіберзалякування. Від'ємні ваги (негативний вплив) вказують, що визначені слова знижують ймовірність належності тексту до певного типу кіберзалякування і є характерними для інших типів [116].

Далі на кроці 2 кортежі з *W* використовуються для формування множини графічних подань інтерпретації результатів виявлення кіберзалякувань *VI* [132]. До множини *VI* входять такі графічні подання: колірне підсвічування у тексті важливих слів w_i та їх ваги β_i для кожного окремого типу кіберзалякування з множини *CBT* [122, 123]. Для цього використовується два способи подання [126]:

1) з використанням абсолютного значення ваг для визначення яскравості кольору при якому не враховується напрямок впливу слів на належність тексту до певного типу кіберзалякування;

2) з використанням значення ваг для визначення кольору та його яскравості, при якому враховується напрямок впливу слів.

Додатково забезпечено побудову діаграми значущості слів для кожного типу кіберзалякувань окремо, а також діаграму середнього значення важливості топ 10 слів для всіх типів кіберзалякувань [127].

Отже, наведений метод інтерпретації результатів виявлення кіберзалякувань дозволяє отримати пояснення щодо прийнятих рішень моделі мультилейблової класифікації текстового контенту. Це дозволяє вирішити проблему зниження довіри до результатів нейромережевих рішень з виявлення кіберзалякувань внаслідок їх низького пояснювання.

2.7.2. Демонстрація роботи методу

Для демонстрації роботи методу використано текстовий зразок з прикладу у пункті 2.6.2 «*Чого ти ще сюди лізеш, стара карга*? *Ніхто не питає думки баби*, *якій давно пора на пенсію*! *Йди краще пиріжки пекти*!».

На першому кроці інтерпретаційна модель визначає важливість слів для кожного з типів кіберзалякувань, наприклад, формуються такі кортежі із слів та їх ваг:

Для вікового {(*"стара", 0.15), ("карга", 0.12), ("баби", 0.11), ("пенсію", 0.09), ("давно", 0.07), ("пора", 0.05), ("пиріжки", -0.02), ("пекти", 0.01), ("думки", -0.04)}.*

Для гендерного *{("баби", 0.14), ("пиріжки", 0.11), ("пекти", 0.09), ("йди", 0.07), ("питає", -0.03)}.*

Для етнічного *{("карга", 0.03), ("пиріжки", -0.02)}.*

Для релігійного *{("пенсію", 0.02), ("пекти", -0.01), ("йди", -0.03)}.*

Для інших типів кіберзалякувань *{("лізеш", 0.08), ("сюди", 0.06), ("чого", 0.05), ("ще", 0.04), ("йди", 0.03), ("краще", 0.02)}.*

На наступному кроці на основі обрахованих ваг відбувається візуальна інтерпретація. Далі наведено приклад інтерпретації з використанням абсолютного значення ваг для визначення яскравості кольору, при якому не враховується напрямок впливу слів на належність тексту до певного типу кіберзалякування. Залежно від величини ваги змінюється яскравість кольору. Чим вона більша – тим яскравіший колір.

Вікове кіберзалякування:

«*Чого ти ще сюди лізеш, стара (0.15) карга (0.12)? Ніхто не питає думки (0.04) баби (0.11), якій давно (0.07) пора (0.05) на пенсію (0.09)! Йди краще пиріжки (0.02) пекти (0.01)!*».

Гендерне кіберзалякування:

«*Чого ти ще сюди лізеш, стара карга? Ніхто не питає (0.03) думки баби (0.14), якій давно пора на пенсію! Йди (0.07) краще пиріжки (0.11) пекти (0.09)!*».

Етнічне кіберзалякування:

«*Чого ти ще сюди лізеш, стара карга (0.03)? Ніхто не питає думки баби, якій давно пора на пенсію! Йди (0.03) краще пиріжки (0.02) пекти!*».

Релігійне кіберзалякування:

«*Чого ти ще сюди лізеш, стара карга? Ніхто не питає думки баби, якій давно пора на пенсію (0.02)! Йди (0.03) краще пиріжки пекти (0.01)!*».

Інший тип кіберзалякування:

«*Чого (0.05) ти ще (0.04) сюди (0.06) лізеш (0.08) , стара карга? Ніхто не питає думки баби, якій давно пора на пенсію! Йди (0.03) краще (0.02) пиріжки пекти!*».

Приклади діаграм значущості слів для кожного типу кіберзалякувань окремо наведено на рис. 2.8.

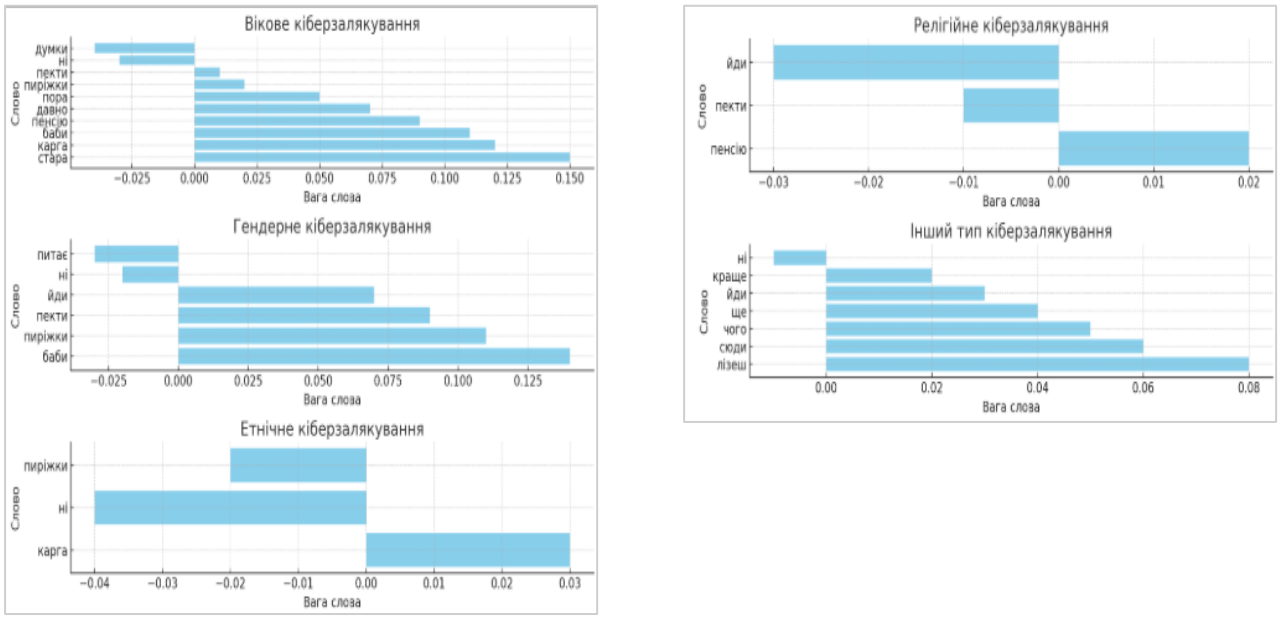


Рис. 2.8 – Приклади діаграм значущості слів для кожного типу кіберзалякувань окремо

Таким чином, запропонований метод інтерпретації результатів виявлення кіберзалякувань надає змогу отримати пояснення щодо рішень моделі мультилейблової класифікації типів кіберзалякувань.

2.8. Висновки до розділу 2

Запропоновано підхід до виявлення та класифікації кіберзалякувань у текстовому контенті, що реалізується як комплексний і поетапний процес, який забезпечує підвищення якості виявлення та класифікації кіберзалякувань.

Розроблено новий метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечує недискримінацію за віковою, етнічною, гендерною і релігійною приналежністю, що дозволило підвищити якість навчання класифікаторів для виявлення кіберзалякувань.

Удосконалено метод виявлення кіберзалякувань у текстовому контенті, який дозволяє аналізувати текстовий контент на загальний рівень прояву кіберзалякувань у ньому, а також виконувати мультилейблову класифікацію, надаючи окремі показники для різних типів кіберзалякувань, що дало можливість підвищити якість виявлення кіберзалякувань

Розроблено новий метод візуальної інтерпретації результатів виявлення та класифікації кіберзалякувань дає змогу здійснювати інтерпретацію результатів для кожного виявленого типу кіберзалякування окремо, що досягається завдяки використанню мультилейблового класифікатора на основі нейромережевої архітектури трансформер та інтерпретаційної моделі. Це дало можливість подавати результати в зрозумілому для користувача вигляді.

Запропонований підхід забезпечує виявлення та класифікацію кіберзалякувань при навчанні нейромереж з урахуванням репрезентативності даних щодо різноманітних соціальних груп, зокрема за віковими, гендерними та релігійними ознаками. Розроблені методи дозволяють виявляти різні типи кіберзалякувань у текстовому контенті та забезпечують пояснення рішень нейромережевої моделі через візуальну інтерпретацію, що підвищує прозорість та довіру до результатів.

РОЗДІЛ 3.
ІНТЕЛЕКТУАЛЬНА ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ КІБЕРЗАЛЯКУВАНЬ

Для експериментального дослідження розроблених у дослідженні методів запропоновано прикладну програмну реалізацію – інтелектуальну інформаційну систему для виявлення та класифікації кіберзалякувань у текстовому контенті. Інформаційна система програмно реалізує наведений в розділі 2 підхід та відповідні йому три розроблені методи (метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, метод виявлення кіберзалякувань у текстовому контенті, метод інтерпретації результатів виявлення кіберзалякувань). У розділі наведено архітектуру інтелектуальної інформаційної системи та спроєктовано її компоненти, описано сформовані зразки даних навчання та тестування нейромережевих моделей, архітектури яких наведено у п. 3.5. Також наведено приклади використання інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті.

3.1. Взаємозв'язок підсистем інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті

Інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань у текстовому контенті складається з чотирьох підсистем. На рис. 3.1 наведено процес перетворення вхідної інформації на вихідну з використанням спроєктованих та розроблених підсистем [137].

Процес виявлення та класифікації кіберзалякувань у текстовому контенті відбувається за допомогою трьох основних підсистем, які реалізують поетапне перетворення вхідних даних на вихідні. Додатково за допомогою підсистеми навчання НМ для етичних аспектів і виявлення та класифікації типів

кіберзалякувань забезпечується отримання моделей для їх використання у підсистемах оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, виявлення та класифікації кіберзалякувань у текстовому контенті.

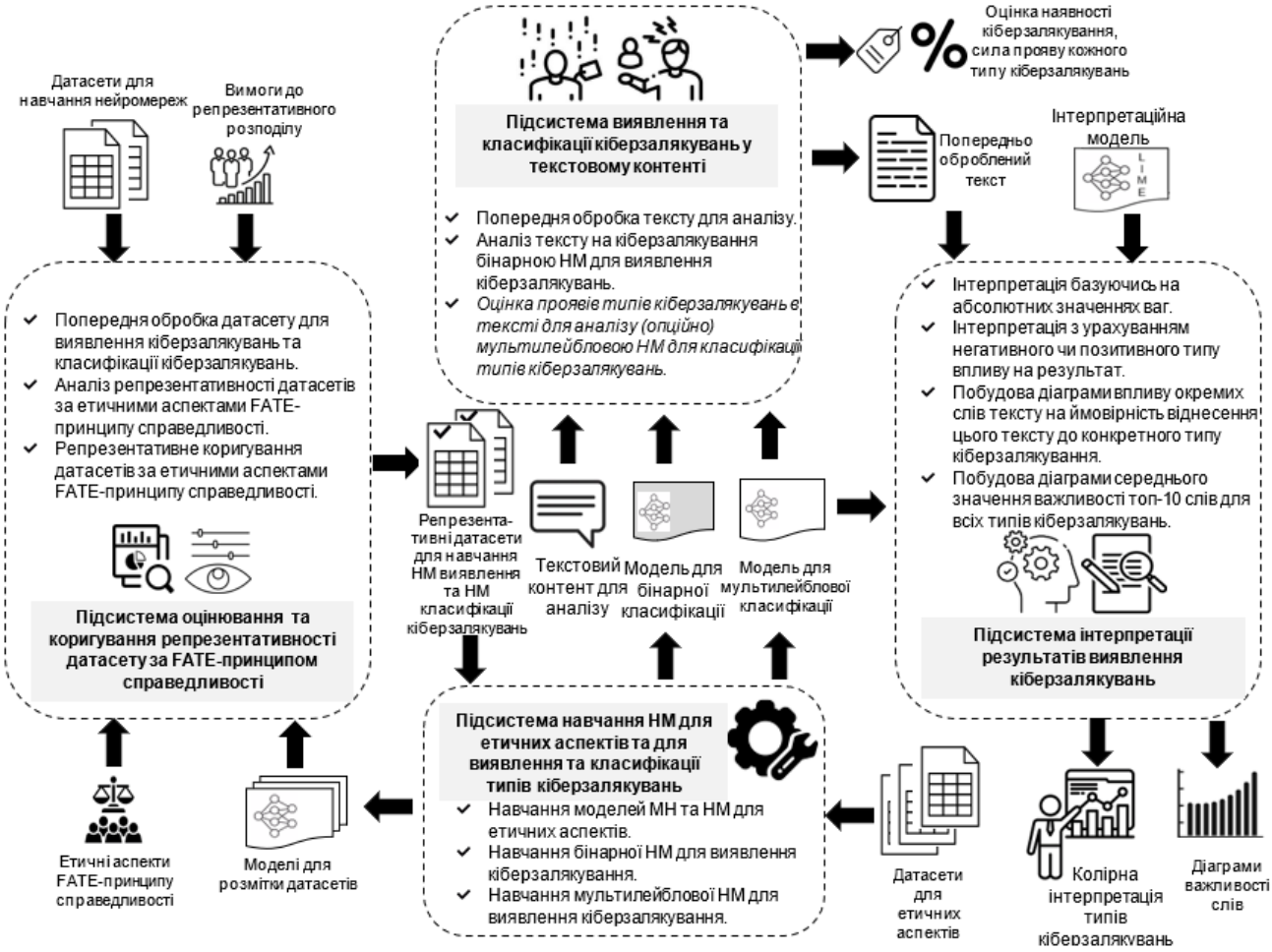


Рис. 3.1 – Процес перетворення вхідної інформації на вихідну
в інтелектуальній інформаційній системі
для виявлення та класифікації кіберзалякувань у текстовому контенті

Підсистема навчання НМ для етичних аспектів, виявлення та класифікації отримує на вхід датасети для етичних аспектів FATE-принципу справедливості, що використовуються для отримання моделей з метою подальшого аналізу на репрезентативність і недискримінаційність датасету для виявлення

кіберзалякувань і датасету для класифікації типів кіберзалякувань. У результаті роботи цієї підсистеми отримуються моделі для кожного етичного аспекту FATE-принципу справедливості, що використовуються у підсистемі оцінювання та коригування репрезентативності датасету. Також після формування репрезентативних і недискримінаційних датасетів у цій підсистемі отримують модель для бінарної класифікації та модель для мультилейблової класифікації, що використовуються в підсистемі виявлення і класифікації кіберзалякувань.

Підсистема оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості отримує на вхід необроблені датасети для навчання нейромереж і вимоги до їхнього репрезентативного та недискримінаційного пропорційного розподілу. В межах цієї підсистеми здійснюється попередня обробка вхідних датасетів, що включає оцінювання їхньої відповідності етичним аспектам FATE-принципу справедливості та, за необхідності, коригування для усунення розбіжностей між фактичними пропорціями датасетів і цільовими. Результатом цього етапу є репрезентативні датасети, які використовуються для подальшого навчання моделі для бінарної класифікації та моделі для мультилейблової класифікації кіберзалякувань у підсистемі виявлення і класифікації кіберзалякувань у текстовому контенті.

Підсистема виявлення та класифікації кіберзалякувань приймає як вхідні дані текстовий контент, що підлягає аналізу, і попередньо навчені моделі нейромереж. На першому кроці здійснюється попередня обробка тексту, після чого бінарна модель нейромережі визначає наявність або відсутність кіберзалякувань. Якщо кіберзалякування виявлено, мультилейблова модель здійснює класифікацію за визначеними типами та обраховує оцінки ймовірності їхніх проявів. На виході цієї підсистеми формується результат, що містить оцінку наявності кіберзалякувань у тексті та оцінки ймовірності їхніх проявів для кожного типу кіберзалякувань.

У підсистемі інтерпретації результатів виявлення кіберзалякувань приймається на вхід оброблений текст з отриманими оцінками ймовірності проявів типів кіберзалякувань та інтерпретаційну модель LIME. В межах цієї підсистеми здійснюється візуальна інтерпретація результатів класифікації, яка надає декілька способів пояснення. Перший ґрунтується на абсолютних значеннях ваг слів для визначення яскравості кольору, при якому не враховується, позитивним чи негативним був вплив слів на належність тексту до певного типу кіберзалякування. Другий спосіб використовує значення ваг для визначення кольору та його яскравості, при якому враховується напрямок впливу слів. Додатково формуються діаграми впливу окремих слів тексту на ймовірність віднесення його до конкретного типу кіберзалякування, а також середнього значення важливості топ 10 слів для всіх типів кіберзалякувань.

Таким чином, процес виявлення та класифікації кіберзалякувань у текстовому контенті включає послідовне перетворення необроблених датасетів у навчені моделі. Подальше застосування цих моделей використовується для етично відповідальних, пояснювальних методів виявлення і класифікації кіберзалякувань у текстовому контенті.

3.2. Функціональні вимоги до інтелектуальної інформаційної системи виявлення та класифікації кіберзалякувань

Проєктування інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті за наведеним потоком перетворення вхідних даних на вихідні передбачає визначення чітких функціональних вимог, що забезпечують її цілісне функціонування, ефективну взаємодію між підсистемами та досягнення поставлених дослідницьких завдань. Функціональні вимоги сформульовані з урахуванням модульної архітектури системи, що включає чотири основні підсистеми: 1) оцінювання

репрезентативності даних; 2) навчання моделей; 3) виявлення кіберзалякувань; 4) пояснення та візуалізації результатів.

Підсистема оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості повинна забезбезпечувати відповідність навчального датасету етичним аспектам [138].

До функціональних вимог належать:

1) попередня обробка датасету, яка включає очищення текстових даних, видалення порожніх або неповних рядків, а також нормалізацію контенту до єдиного формату;

2) оцінка відповідності реального розподілу класів у датасеті до цільових пропорцій;

3) можливість балансування диспропорцій у датасеті за допомогою видалення та аугментації даних у класах;

4) збереження відкоригованого датасету для подальшого використання в процесах навчання моделей для завдань виявлення та класифікації кіберзалякувань у текстовому контенті.

Підсистема навчання НМ для етичних аспектів, виявлення і класифікації повинна забезпечувати навчання та валідацію нейромережевих моделей, які використовуються для оцінювання репрезентативності даних, виявлення кіберзалякувань і класифікації їх типів.

Основні функціональні вимоги включають:

1) завантаження датасетів, підготовлених підсистемою репрезентативності датасету за FATE-принципом справедливості та їх первинна обробка (очищення, токенізація, векторизація);

2) реалізація процесів навчання моделей, зокрема:

– моделей бінарної класифікації для виявлення наявності кіберзалякувань;

– моделей мультилейблової класифікації для визначення типів кіберзалякувань;

– моделей етичної етичних аспектів для оцінювання репрезентативності датасетів;

3) проведення валідації натренованих моделей з обчисленням метрик якості (точність, повнота, влучність, F1-міра);

4) формування звітів за результатами навчання та збереження моделей для використання іншими підсистемами системи.

Підсистема виявлення та класифікації кіберзалякувань повинна реалізовувати безпосередню класифікацію текстових повідомлень за ознаками кіберзалякувань. До функціональних вимог належать:

1) прийом на вхід текстових повідомлень користувача та їх попередня обробка;

2) здійснення бінарної класифікації вхідного тексту за наявністю ознак кіберзалякування на основі моделі BiLSTM;

3) у разі виявлення кіберзалякування – застосування моделі BERT для мультилейблової класифікації типів кіберзалякуваннь;

4) формування результату класифікації у вигляді оцінки наявності кіберзалякувань та їх типів.

Підсистема інтерпретації результатів виявлення кіберзалякувань повинна забезпечувати пояснювання результатів класифікації. Вимоги до цієї підсистеми:

1) генерація локальних пояснень до рішень моделі за допомогою алгоритму LIME, що дозволяє встановити, які саме слова у тексті вплинули на результат;

2) підсвічування важливих слів у тексті із зазначенням ваги їхньої важливості для виявленого типу кіберзалякування;

3) відображення візуальної інформації у вигляді діаграм важливості, що дозволяють користувачу оцінити, наскільки окремі слова впливали на результат класифікації як окремого типу кіберзаялкування, так і загалом;

4) можливість обрання методу візуалізації пояснення прийнятих нейромережевою моделлю рішень: за абсолютним значенням важливості або з урахуванням напрямку впливу.

Розробка функціональних вимог до інтелектуальної інформаційної системи виявлення та класифікації кіберзалякувань є важливим етапом, що визначає подальшу архітектуру системи та логіку її функціонування. Визначені вимоги охоплюють усі етапи обробки вхідного текстового контенту – від попереднього оцінювання репрезентативності датасету за етичними аспектами до безпосередньої класифікації кіберзалякувань і інтерпретації результатів у вхідному зразку.

3.3. Архітектура інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті

На основі сформульованих функціональних вимог було спроєктовано архітектуру інтелектуальної інформаційної системи, що наведена на рис. 3.2. Відповідно до цієї архітектури було розроблено програмну реалізацію системи, що забезпечує експериментальне впровадження й дослідження ефективності запропонованих методів виявлення кіберзалякувань з урахуванням принципів етичності, репрезентативності та пояснювання.

Розробка інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті та експериментальне дослідження запропонованих у дисертаційній роботі методів проводилися на 64-розрядній операційній системі Windows 11 Home. Для реалізації програмного забезпечення використано Python версії 3.12 та фреймворк Flask. Для створення графічного інтерфейсу та реалізації моделей, таких як BERT, BiLSTM, SVM, Random Forest, LSTM, GRU, RoBERTa, а також методу інтерпретації LIME використано бібліотеки PyQt6 v6.6.0, Transformers v4.35.1, TensorFlow v2.14.0,

PyTorch v2.1.0, NumPy v1.26.0, Pandas v2.1.2, Keras v2.14.0, Scikit-learn v1.5.2, LIME v0.2.0.1, Matplotlib v3.8.1, Seaborn v0.12.2.

Параметри ПК, на якому проводилась розробка інформаційної системи та експерименти:

– процесор AMD Ryzen 5 5500U з 6 ядрами та 12 потоками, базова частота – 2.10 ГГц

– оперативна пам’ять – 16 ГБ;

– графічний процесор Radeon Graphics.

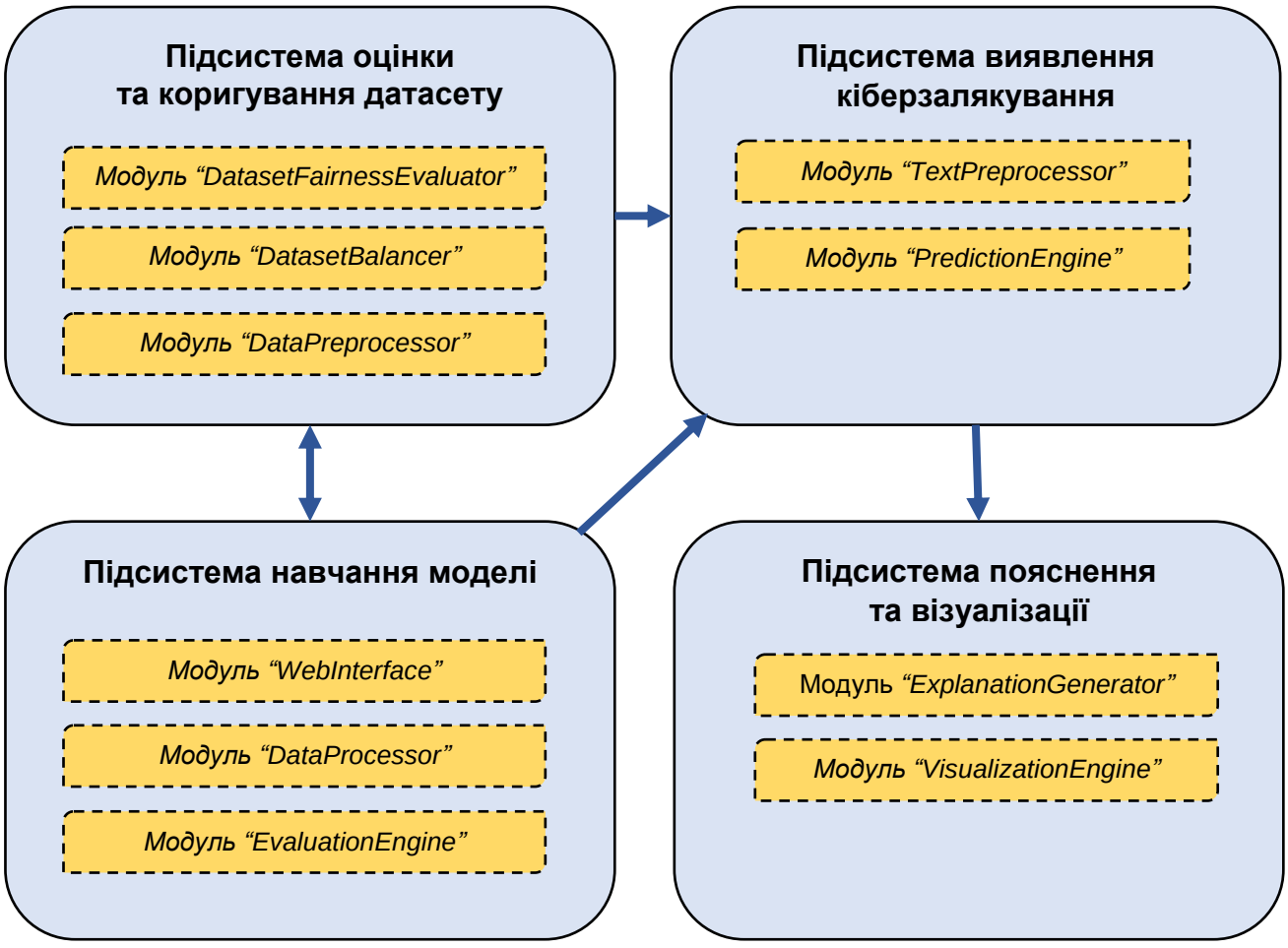


Рис. 3.2 – Архітектура інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті

Архітектура інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті наведена на рис. 3.2.

Підсистема оцінювання та коригування репрезентативності датасету («Dataset Evaluation and Adjustment Subsystem») забезпечує аналіз та коригування датасету для забезпечення його репрезентативності за FATE-принципом справедливості. Модуль «DataPreprocessor» відповідає за попередню обробку даних, включаючи очищення текстів, видалення порожніх рядків, що є необхідною умовою для подальшого навчання моделі. Модуль «DatasetFairnessEvaluator» виконує перевірку розбіжності між фактичними пропорціями у завантаженому датасеті та цільовими. «DatasetBalancer» дозволяє здійснювати коригування репрезентативності шляхом балансування даних за допомогою SMOTE-балансування.

Підсистема виявлення кіберзалякувань у текстовому контенті («Cyberbullying Detection Subsystem») є частиною системи, що забезпечує виявлення та класифікацію кіберзалякувань у текстах. Модуль «TextPreprocessor» займається очищенням, нормалізацією та токенізацією вхідного тексту, що дозволяє підготувати його для аналізу нейромережевими моделями. Модуль «PredictionEngine»” реалізовує бінарну класифікацію за допомогою HM BiLSTM та мультикласової класифікації типів кіберзалякувань за допомогою BERT.

Підсистема інтерпретації результатів виявлення класифікації кіберзалякувань («Explanation and Visualization Subsystem») спрямована на забезпечення прозорості та пояснювання роботи моделі. Модуль «ExplanationGenerator» використовує інтерпретаційну модель LIME для створення локальних пояснень, які показують, які саме слова або фрази вплинули на рішення моделі BERT при мультилейбловій класифікації типів кіберзалякувань. Модуль «VisualizationEngine» відповідає за візуалізацію цих пояснень у зручній для користувача формі, підсвічуючи слова з відповідними значеннями ваг та виводячи

їх значення. Також наведений модуль відповідає за побудову діаграм важливості слів.

Підсистема навчання моделей («Model Training Subsystem») відповідає за навчання та валідацію моделей для етичних аспектів, виявлення та класифікації кіберзалякувань у текстовому контенті. Вона приймає на вхід датасети для етичних аспектів, а також репрезентативні датасети для виявлення та класифікації кіберзалякувань з «Dataset Evaluation and Adjustment Subsystem». Модуль «DataProcessor» виконує попередню обробку вхідних текстових даних, що включає очищення, токенізацію та перетворення тексту у векторне представлення. Модуль «ModelTrainer» відповідає за процес навчання моделей. Навчання здійснюється як для бінарної класифікації моделей для етичних аспектів, що використовуються у підсистемі «Dataset Evaluation and Adjustment Subsystem», так і моделей для виявлення та класифікації у підсистемі «Cyberbullying Detection Subsystem».

Після навчання моделі відбувається процес валідації. Модуль «EvaluationEngine» обчислює метрики, такі як точність, повнота, влучність та F_1-міра.

Вихідними даними підсистеми є навчені моделі, що використовуються у «Dataset Evaluation and Adjustment Subsystem» для аналізу датасетів на репрезентативність і неупередженість, а також у підсистемі «Cyberbullying Detection Subsystem» для подальшого використання у виявленні та класифікації кіберзалякувань у текстах. Також система створює звіти з наведеними вище метриками.

Відповідно до описаної архітектури інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті розроблено далі наведено опис функцій та діаграму класів (рис. 3.3).

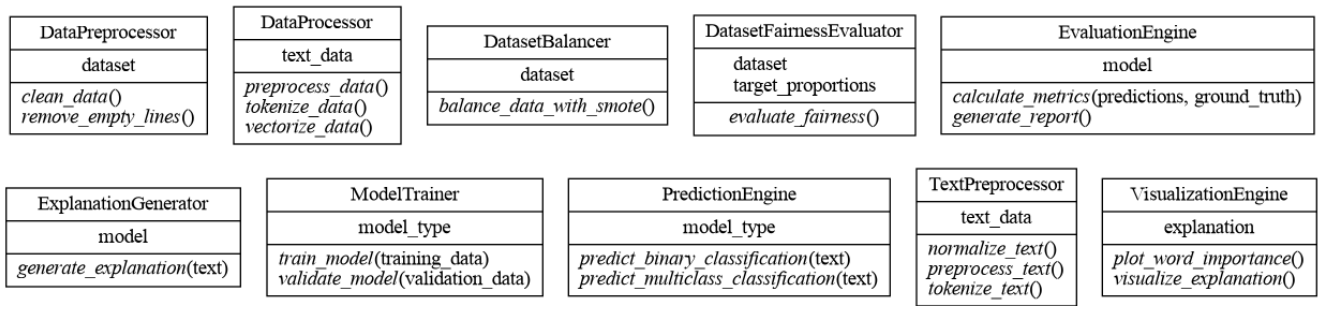


Рис. 3.3 – Діаграма класів розробленої інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті

Підсистема оцінки та коригування датасету.

1. Модуль «DataPreprocessor» та його функції:
   - clean_data() очищує дані від шумів, непотрібних символів, що можуть впливати на якість моделі;
   - remove_empty_lines() видаляє порожні рядки з датасету.
2. Модуль «DatasetFairnessEvaluator» та його функції:
   - evaluate_fairness() порівнює фактичний розподіл зразків у датасеті із заданими цільовими пропорціями.
3. Модуль «DatasetBalancer» та його функції:
   - balance_data_with_smote() виконує балансування класів у датасеті за допомогою алгоритму SMOTE для генерації синтетичних прикладів менш представлених класів, а також виконує видалення надмірно представлених даних у класах.

Підсистема виявлення кіберзалякувань у текстовому контенті.

1. Модуль «TextPreprocessor» та його функції:
   - normalize_text() нормалізує текст (наприклад, зводить до нижнього регістру, замінює скорочення);
   - preprocess_text() попередньо обробляє текст (видаляє спецсимволи, знаки пунктуації);

- tokenize_text() розбиває текст на токени (слова або підслова).

2. Модуль «PredictionEngine» та його функції:
   - predict_binary_classification(text) реалізує бінарну класифікацію тексту (кіберзалякування/некіберзалякування), зокрема на основі BiLSTM;
   - predict_multiclass_classification(text) виконує мультилейблову класифікацію типів кіберзалякування за допомогою BERT.

Підсистема пояснення та візуалізації.

1. Модуль «ExplanationGenerator» та його функції:
   - generate_explanation(text) створює локальні пояснення класифікаційного рішення на основі моделі LIME.
2. Модуль «VisualizationEngine» та його функції:
   - plot_word_importance() будує діаграму, що відображає важливість слів у прийнятті рішення;
   - visualize_explanation() візуалізує пояснення для тексту – підсвічує ключові слова та їхні ваги.

Підсистема навчання моделі.

1. Модуль «DataProcessor» та його функції:
   - preprocess_data() виконує початкову обробку тексту (загальна функція для очищення);
   - tokenize_data() здійснює токенізацію для подальшої обробки моделлю;
   - vectorize_data() перетворює текст на числове представлення (вектори) для подачі в модель.
2. Модуль «ModelTrainer» та його функції:
   - train_model(training_data) здійснює навчання моделі на вхідних навчальних даних;
   - validate_model(validation_data) оцінює якість моделі на валідаційних даних для перевірки узагальнюючої здатності.

3. Модуль «EvaluationEngine» та його функції:

- calculate_metrics(predictions, ground_truth) обчислює основні метрики (точність, повнота, влучність, F1-міра);
- generate_report() формує звіт про точність моделі на основі обчислених метрик.

Отже, для прикладної дослідницької програмної реалізації розроблених у дослідженні методів, спроєктовано архітектуру інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті, що складається з чотирьох підсистем і здатна забезпечити можливість експериментального дослідження розроблених методів.

3.4. Формування датасетів для навчання та валідування моделей машинного навчання

Для експериментального навчання моделей машинного навчання виникла необхідність використання ряду датасетів. Для навчання моделей, що виконуватимуть розмітку датасетів для виявлення кіберзалякувань, сформовано три датасети, що відповідають трьом етичним аспектам FATE-принципу справедливості: віковому, гендерному та релігійному. Для виявлення та класифікації кіберзалякувань також необхідно сформувати відповідні датасети для навчання моделей. Перший датасет використовуватиметься для навчання бінарного класифікатора, а другий – для мультилейблового. Далі наведено детальну інформацію про використані датасети.

Формування датасетів для оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості. Для навчання нейромережевих моделей методу оцінювання і коригування репрезентативності датасету за FATE-принципом справедливості, призначених

для розмітки вхідного датасету, використано датасети, що відображають три етичні аспекти принципу справедливості: гендер, вік і релігію.

Усі датасети, використані в дослідженні, були перекладені з оригінальних мов на українську за допомогою бібліотеки Googletrans 4.0.2 [140]. Після завершення перекладу отримані дані використовувалися для навчання нейромережевих моделей, що дозволило їм класифікувати текстові зразки українською мовою.

Зокрема, для визначення статі автора повідомлення застосовано англомовний датасет «Tweet Files for Gender Guessing» [139], який містить 34146 текстових записів, рівномірно розподілених між двома класами: чоловіки та жінки (по 17073 записи у кожному).

Для навчання класифікатора з урахуванням релігійного аспекту використано вибірку, створену на основі англомовного датасету «CyberBullying Detection Dataset» [43], що містить 20109 тестових зразків. З датасету було використано такі класи як christian (1163 зразків), buddhism (788 зразків), muslim (2559 зразків), jew (1950 зразків).

Крім того, італійськомовний датасет «TAG-it Dataset Distribution» [141] було перекладено українською мовою та застосовано для репрезентативного відображення вікової характеристики текстових повідомлень. Він містить 21948 зразків, розподілених за віковими категоріями: 0–19 (3126 зразків), 20–29 (6568 зразків), 30–39 (6568 зразків), 40–49 (4178 зразків), 50–100 років (3865 зразків).

Оскільки наведені датасети містили нерівномірно представлені класи, що впливатиме на якість навчання моделей машинного навчання, усі класи було збалансовано. Остаточний розподіл даних у навчальних датасетах, що використовувалися для навчання моделей для розмітки за етичними аспектами, наведено на рисунку 3.3.

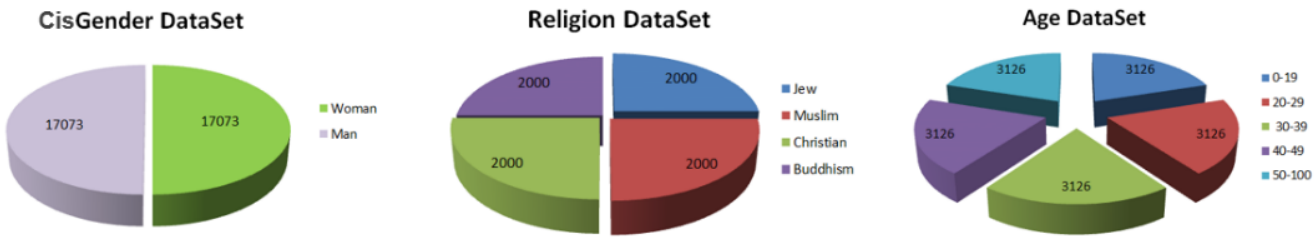


Рис. 3.3 – Класи та кількість зразків у датасетах
для навчання моделей за етичними аспектами

У результаті роботи зі створення навчальних датасетів, отримано збалансовані за кількістю текстових повідомлень у класах датасети. Такі збалансовані датасети дозволять коректно оцінювати репрезентативність текстових датасетів для виявлення та класифікації кіберзалякувань.

Формування датасетів для навчання та валідування моделей виявлення кіберзалякувань у текстовому контенті. Для навчання класифікаторів та оцінки ефективності методу виявлення кіберзалякувань було використано два датасети: «Cyberbullying Tweets» [141] для задачі бінарної класифікації та «Cyberbullying Classification» [42] для мультилейблової класифікації.

Датасет «Cyberbullying Tweets» містить 11100 унікальних текстових записів, поділених на дві категорії: повідомлення з ознаками кіберзалякувань та ті, що їх не містять. Обидва класи у вибірці представлені рівномірно по 5 550 текстових зразків у кожному.

Датасет «Cyberbullying Classification» включає 46017 унікальних записів, розподілених за категоріями: Age, Ethnicity, Gender, Religion, Other type of cyberbullying та Not cyberbullying. Детальний розподіл обсягів кожного класу представлено на рис. 3.4.

Аналіз розподілу класів у датасеті (рис. 3.4) свідчить про їхню нерівномірність, що впливатиме на точність навчання моделей машинного навчання. Для вирішення цієї проблеми було застосовано SMOTE-балансування,

яке дозволило синтетично збалансувати вибірку, довівши кількість зразків у кожному класі до 7998. Для навчання моделі BERT у задачі мультилейблової класифікації не використовувався клас «Not cyberbullying».

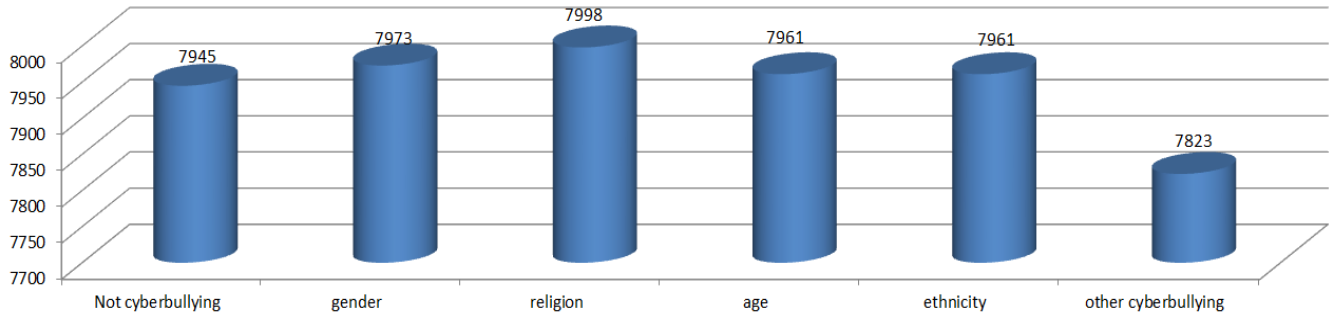


Рис. 3.4 – Розподіл повідомлень датасету за класами

Варто зазначити, що обидва датасети були оцінені та скориговані за допомогою методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, після чого використані для навчання нейромережевої моделі для виявлення кіберзалякувань та моделі для класифікації типів кіберзалякувань.

3.5. Архітектури моделей машинного навчання

Архітектури моделей машинного навчання для оцінювання репрезентативності датасетів за етичними аспектами. Для кожного етичного аспекту, що оцінюватимуть репрезентативність вхідного датасету на Кроці 2 методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості обрано такі нейромережеві моделі глибокого навчання, як LSTM – для гендерного етичного аспекту (рис. 3.5) та BERT – для релігійного етичного аспекту (рис. 3.6), а також класифікатор SVM – для вікового етичного аспекту [145].

Моделі обрані з огляду на проведені експериментальні дослідження різних моделей машинного та глибокого навчання у розділі 4. Отримані результати підтверджують здатність LSTM та BERT класифікувати текстові зразки у датасетах для гендерного та релігійного аспектів, відповідно.

Наведена на рис. 3.5 архітектура нейронної мережі LSTM містить чотири основні шари. Першим шаром є шар embedding, який перетворює вхідні цілі числа, що представляють токени, у векторні представлення фіксованої розмірності. Вхідні дані мають форму (None, 512), що вказує на змінну кількість зразків, кожен з яких містить 512 токенів. Вихідний тензор має форму (None, 512, 128), що означає, що кожен із 512 токенів представлений у 128-вимірному просторі ознак.

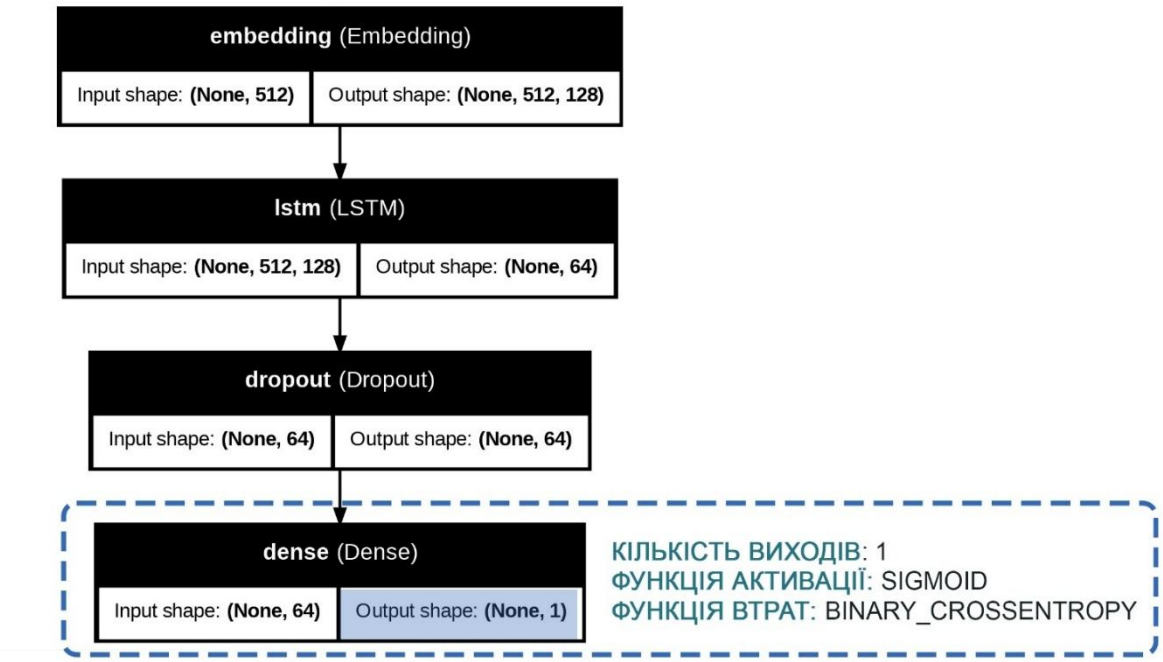


Рис. 3.5 – Архітектура нейронної мережі LSTM для розмітки елементів датасету за гендерним етичним аспектом

Наступним шаром є рекурентний LSTM, який отримує вхідні дані у вигляді тензора розмірності (None, 512, 128). Цей шар обробляє послідовність довжиною

512, використовуючи прихований стан розмірністю 128. Після обробки він повертає тензор форми (None, 64), що вказує на те, що він передає лише останній прихований стан кожної послідовності до наступного, оскільки `return_sequences=False`.

Далі використовується шар dropout, який отримує вхідні дані розмірності (None, 64) і виконує регуляризацію шляхом випадкового занулення деяких нейронів під час навчання, що дозволяє зменшити перенавчання моделі та покращити її здатність до узагальнення. Вихідні дані зберігають ту саму розмірність (None, 64).

Завершальним шаром є dense шар, який отримує вхідні тензори форми (None, 64) і застосовує лінійну або нелінійну активацію для перетворення вектора ознак у вихідне значення форми (None, 1).

Наведена на рис. 3.6 архітектура нейронної мережі BERT для релігійного аспекту складається з 12 шарів. Вхідні дані складаються з двох тензорів: input_ids, що містить закодовані токени, та attention_mask, яка визначає, які токени слід враховувати під час обробки. Обидва вхідні тензори мають форму (None, 512), що означає змінну кількість зразків, де кожен текст представлений послідовністю з 512 токенів.

На наступному етапі застосовується шар BERT_Embedding, який є лямбда-функцією для отримання векторного представлення тексту розмірністю 768. Вихідний тензор має форму (None, 768), що відповідає прихованому стану вихідного шару BERT.

Подальша обробка здійснюється dense-шарами, нормалізацією мініпакетів (BatchNormalization) та dropout-шарами. Перший повнозв’язний шар dense1 отримує на вхід вектор розмірності 768 і перетворює його на вектор розміром 256. Далі застосовується batch_norm1, який нормалізує вихідні значення для покращення стабільності навчання, і dropout1, що випадковим чином зануляє певні нейрони для зменшення перенавчання.

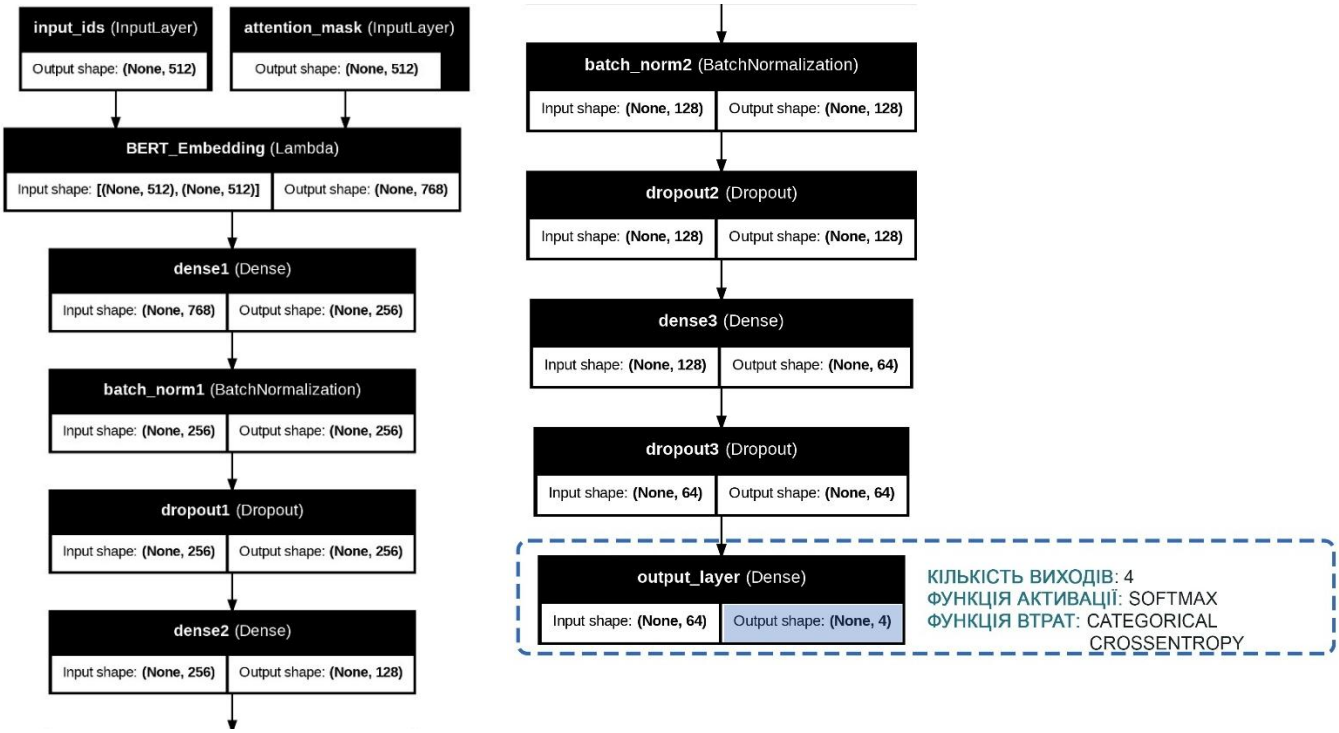


Рис. 3.6 – Архітектура нейронної мережі BERT
для розмітки елементів датасету за релігійним етичним аспектом

Наступний повнозв’язний шар dense2 зменшує розмірність до 128, після чого використовуються batch_norm2 і dropout2 для покращення узагальнюючої здатності моделі. Потім застосовується dense3, який перетворює вхідний вектор на 64-вимірний простір ознак, після чого додається dropout3.

Останній шар output_layer є повнозв’язним шаром, що отримує вхідний тензор розмірності (None, 64) і формує вихідний вектор з 4 компонентами, що вказує на кількість класів у задачі багатокласової класифікації. У вихідному шарі використовується функція активації softmax та функція втрат categorical crossentropy, так як модель застосовується для багатокласової класифікації.

Отже, для оцінювання репрезентативності вхідного датасету на Кроці 2 методу оцінювання та коригування репрезентативності за FATE-принципом справедливості були обрані нейромережеві моделі машинного і глибокого

навчання. Зокрема, для аналізу гендерного етичного аспекту застосовано модель LSTM, для релігійного етичного аспекту – модель BERT, для вікового етичного аспекту – класифікатор SVM.

Архітектури нейромережевих моделей глибокого навчання для виявлення та класифікації типів кіберзалякувань. Для виявлення кіберзалякувань на Кроці 2 методу виявлення та класифікації типів кіберзалякувань навчено нейромережеву модель BiLSTM для бінарної класифікації кіберзалякувань (рис. 3.7) та нейромережеву модель BERT – для мультилейблової класифікації типів кіберзалякувань.

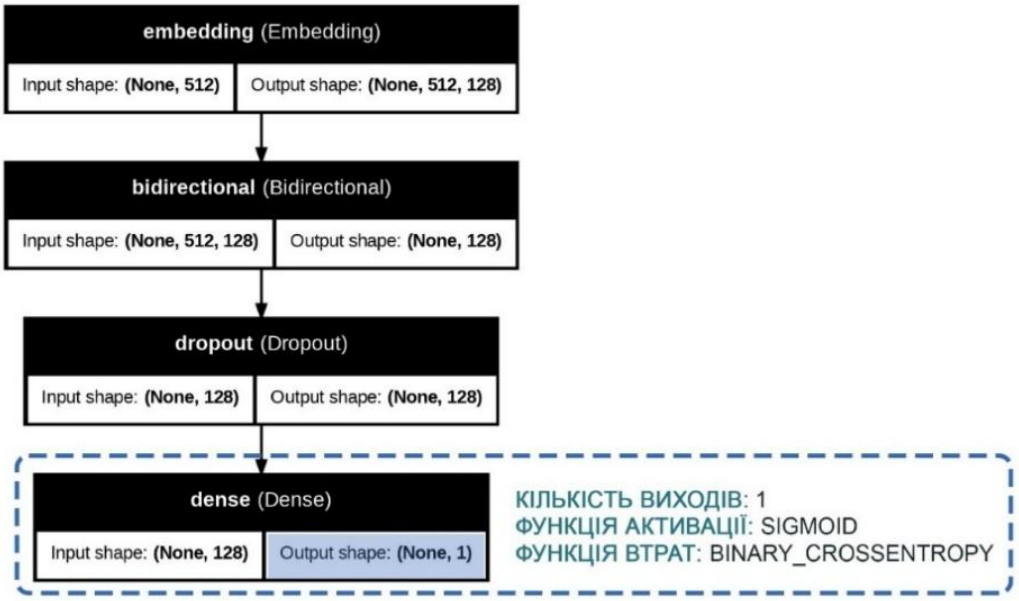


Рис. 3.7 – Архітектура нейронної мережі BiLSTM
для бінарної класифікації кіберзалякувань у текстовому контенті

Використання архітектури BiLSTM для бінарної класифікації кіберзалякувань і моделі BERT для мультилейблової класифікації забезпечує вищу точність і стабільність результатів порівняно з іншими алгоритмами машинного навчання, оскільки BiLSTM ефективно моделює послідовності, враховуючи

контекст у двох напрямах, а BERT використовує механізм самоуваги, що дозволяє краще розуміти семантичні залежності в тексті.

Для підтвердження цієї гіпотези експериментально порівняно різні моделі машинного та глибокого навчання у розділі 4. Отримані результати експериментів підтверджують, що BiLSTM більш точно виявляє кіберзалякування, тоді як BERT забезпечує точнішу класифікацію їх типів.

На рис. 3.7 наведена архітектура BiLSTM для бінарної класифікації, що починається зі embedding-шару, який перетворює вхідну послідовність із 512 токенів на вектори розмірності 128 для кожного токена. Це дозволяє представити текст у вигляді числових векторів, які можуть бути оброблені нейронною мережею.

Наступним є двонаправлений LSTM, який обробляє послідовність у двох напрямках – вперед і назад, що дозволяє моделі захоплювати як попередній, так і наступний контекст для кожного слова у текстовому зразку, що аналізується на наявність кіберзалякування. На виході цей шар формує вектор представлення розмірності 128, що містить зведену інформацію про всю послідовність.

Після цього застосовується шар dropout, який випадково занулює частину нейронів, що необхідно для зменшення ризику перенавчання та покращення узагальнюючої здатності моделі. Вхід і вихід цього шару мають однакову розмірність – 128.

Останнім є повнозв’язний вихідний dense-шар, який отримує вхідні дані розмірності 128 і формує підсумковий прогноз. Він використовує функцію активації softmax для обчислення ймовірностей класів «кіберзалякування» та «некіберзалякування», що дозволяє отримати значення в діапазоні від 0 до 1.

Для мультилейблової класифікації типів кіберзалякувань використовується архітектура нейронної мережі BERT, що має 5 виходів для кожного типу кіберзалякувань (вікове, гендерне, етнічне, релігійне, інші типи кіберзалякувань), у вихідному шарі застосовується sigmoid-активація та функція втрат binary crossentrорує (рис. 3.8).

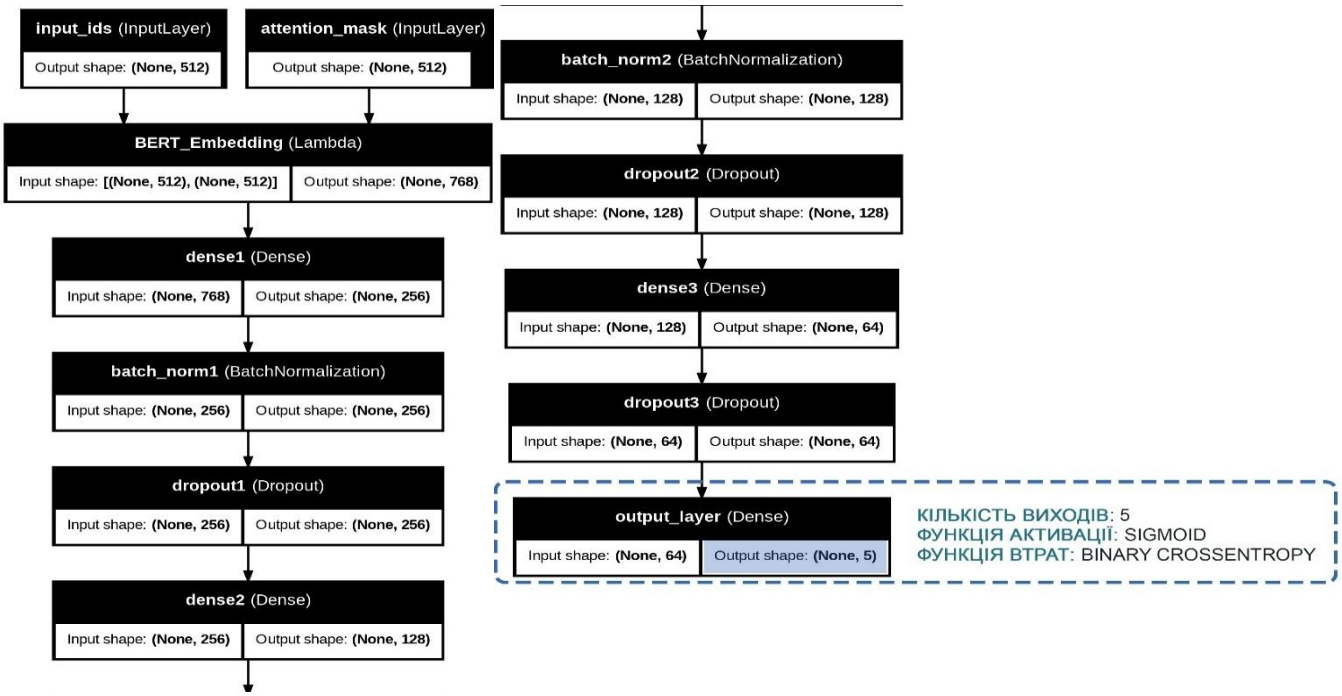


Рис. 3.8 – Архітектура нейронної мережі BERT
мультилейблової класифікації типів кіберзалякувань

Отже, для виявлення та класифікації кіберзалякувань у текстовому контенті використовується дві нейромережеві моделі. Для виявлення (бінарної класифікації) кіберзалякувань використовується архітектура BiLSTM, а для класифікації типів кіберзалякувань (мультилейблова класифікація) використовується архітектура BERT, яка має останній вихідний шар на 5 виходів для кожного типу і використовується sigmoid-активація та функція втрат binary crossentropy.

3.6. Особливості реалізації інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань

Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості було реалізовано у вигляді окремої підсистеми

інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту.

На рис. 3.9 – 3.11 наведено приклади результатів класифікації за етичними аспектами FATE-принципу справедливості.

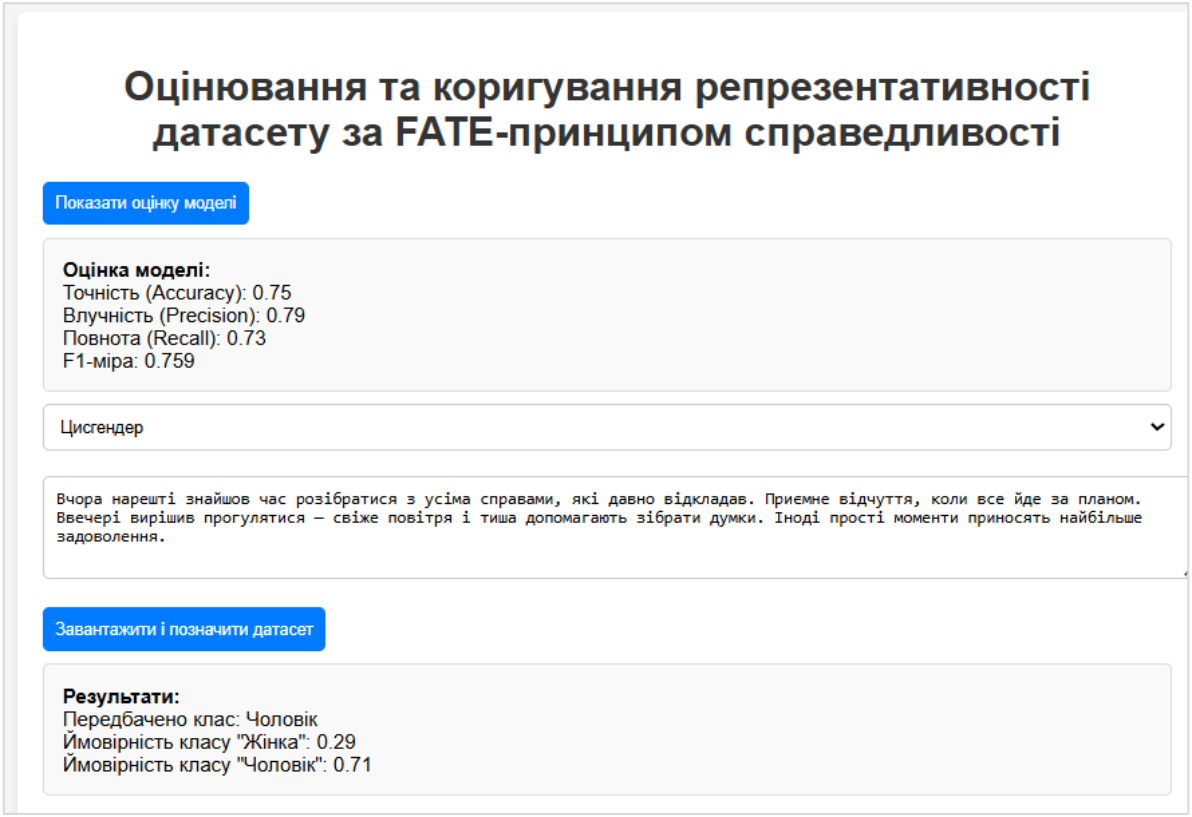


Рис. 3.9 – Програмна реалізація класифікації за гендерною ознакою

Інтерфейс підсистеми містить таблицю з текстовими зразками та їхніми мітками. У таблиці вказано сам текст, тип кіберзалякування, стать користувача, релігію та вік. Ці дані ілюструють мітки для текстових зразків датасету за етичними аспектами FATE-принципу справедливості, що у подальшому буде використано для приведення датасету до репрезентативного вигляду на кроці коригування. Таким чином, підсистема дозволяє не лише оцінювати репрезентативність датасетів, але й коригувати їх відповідно до цільових пропорцій. У підсистемі можна обрати конкретну ознаку для аналізу, наприклад,

вік, і завантажити відповідний датасет. Після аналізу зразка тексту система видає ймовірності приналежності до конкретної вікової групи.

На рис. 3.9 модель продемонструвала точність – 75 %, влучність – 79 %, повноту – 73 % і F1-міру – 75,9 %. При аналізі введеного тексту «*Вчора нарешті знайшов час розібратися з усіма справами, які давно відкладав. Приємне відчуття, коли все йде за планом. Ввечері вирішив прогулятися – свіже повітря і тиша допомагають зібрати думки. Іноді прості моменти приносять найбільше задоволення*», модель визначила, що він із ймовірністю 71 % належить чоловікові та 29 % – жінці.

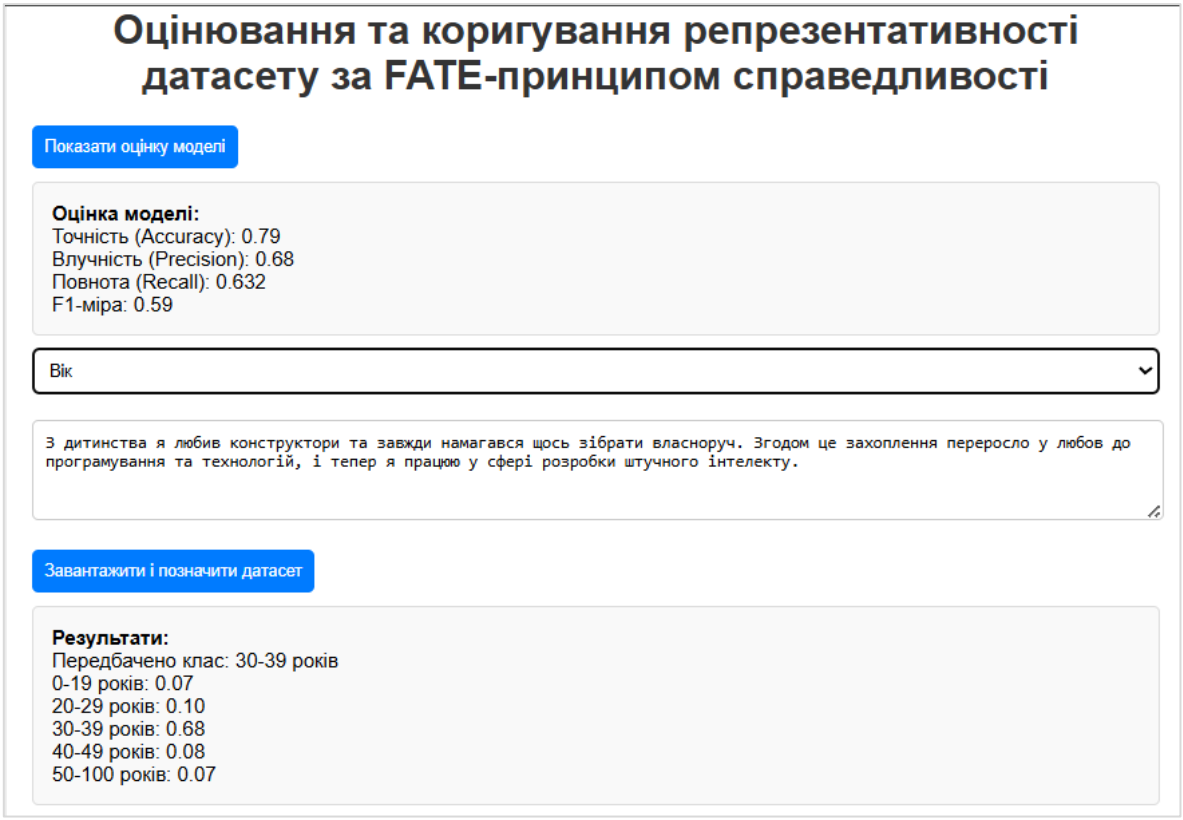


Рис. 3.10 – Програмна реалізація класифікації за віковою ознакою

На рис. 3.10 модель продемонструвала точність – 79 %, влучність – 68 %, повноту – 63,2 % і F1-міру – 65 %.

При аналізі введеного тексту «*З дитинства я любив конструктори та завжди намагався щось зібрати власноруч. Згодом це захоплення переросло у*

любов до програмування та технологій, і тепер я працюю у сфері розробки штучного інтелекту», модель визначила, що він із ймовірністю 7 % належить до вікової групи 0–19 років, 10 % – до групи 20–29 років, 68 % – до групи 30–39 років, 8 % – до групи 40–49 років і 7 % – до групи 50–100 років.

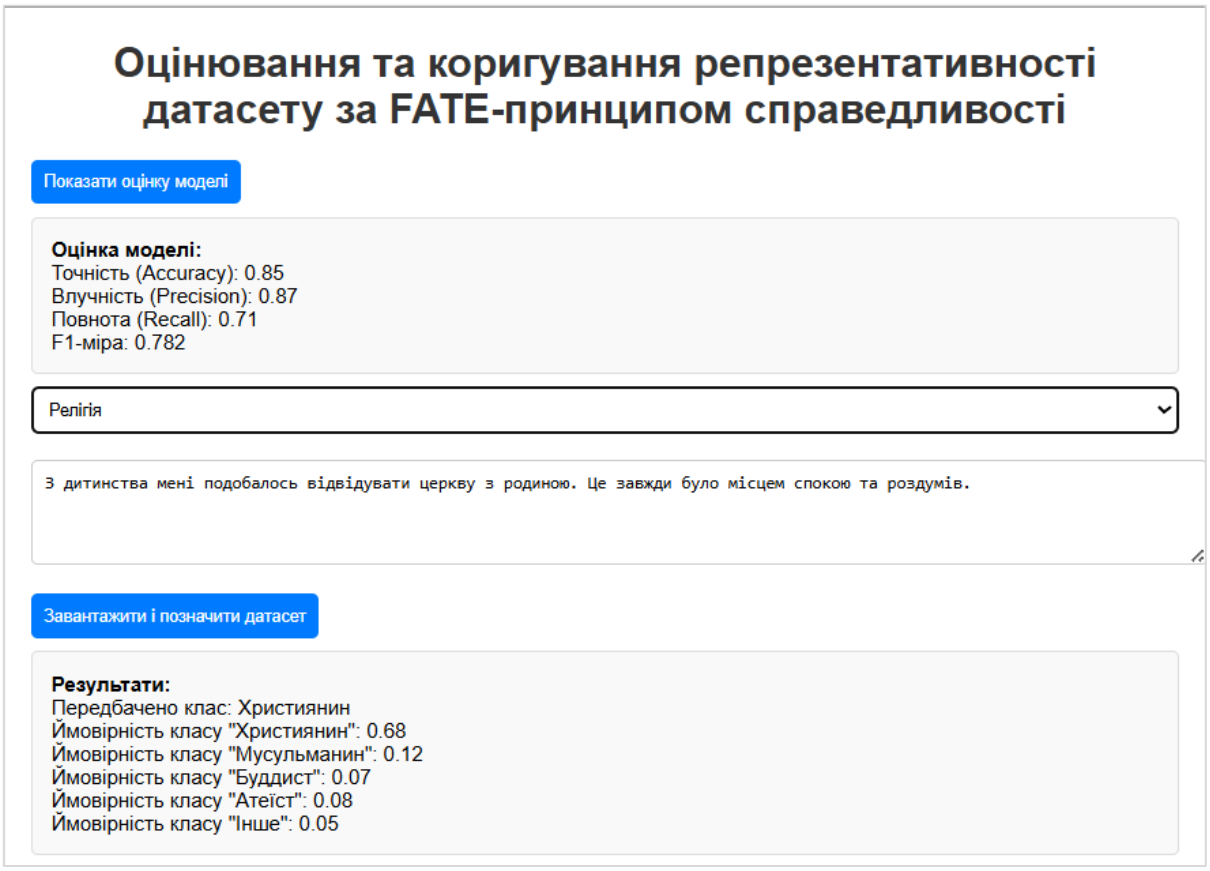


Рис. 3.11 – Програмна реалізація класифікації за релігійною ознакою

На рис. 3.11 модель продемонструвала точність – 85 %, влучність – 87 %, повноту – 71 % і F1-міру – 78,2. При аналізі введеного тексту «*З дитинства мені подобалось відвідувати церкву з родиною. Це завжди було місцем спокою та роздумів*», модель визначила, що він із ймовірністю 68 % належить до класу «Християнин», 12 % – «Мусульманин», 7 % – «Буддист», 8 % – «Атеїст» і 5 % – «Інше».

Також було реалізовано наочне подання міток текстових зразків датасету щодо кожного етичного аспекту (рис. 3.12).

Оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

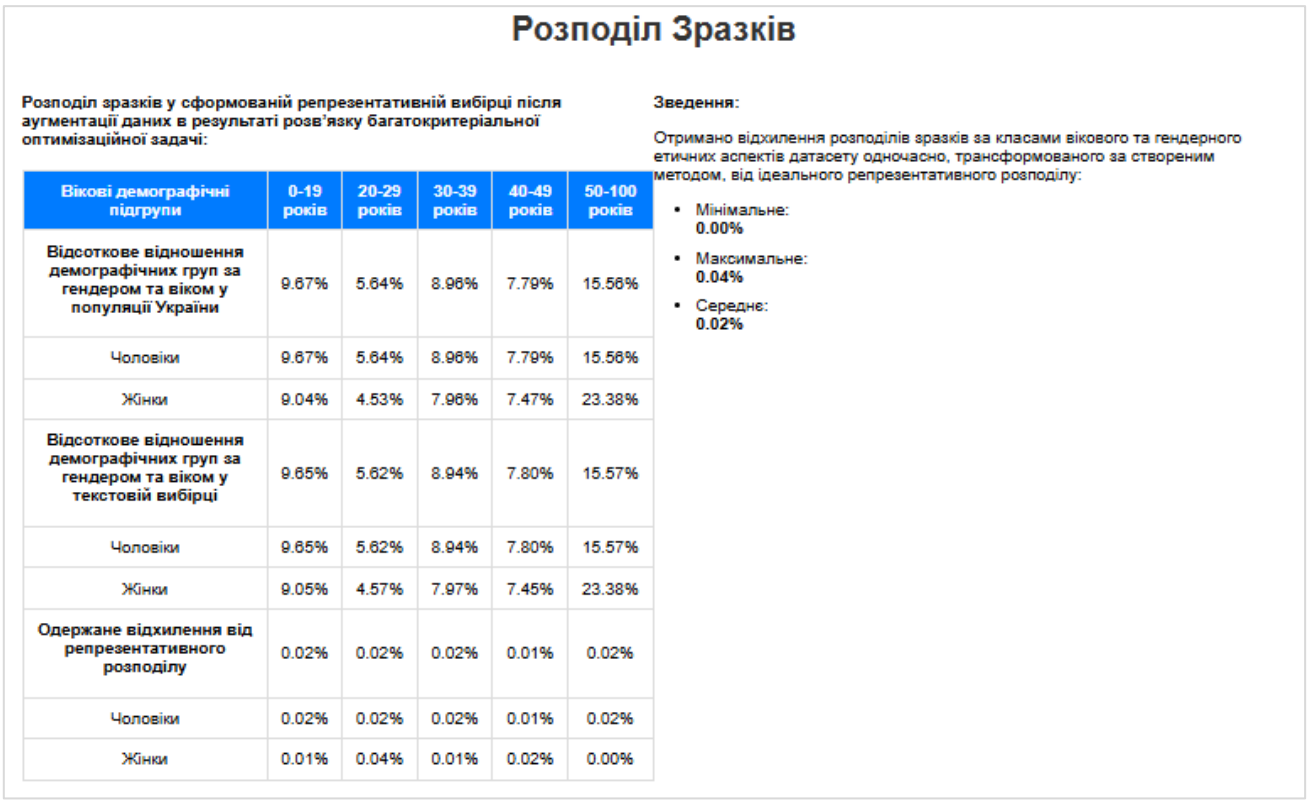
Мітки текстових зразків датасету:

Текстовий зразок	Цисгендер	Релігія	Вік
Ти знову все порушив! Як можна таке робити? Потрібно бути більш відповідальним, а то вже не знаєш, як все полагодити...	Чоловік	Атеїст	20-29
Ну ти і йолоп! Не розумієш елементарного, як тобі взагалі по життю дається щось? Це ж неважко було зробити правильно.	Жінка	Мусульманин	30-39
Ти якесь нерозумне, намагаєшся втягнути в це, а нічого не виходить. Вже ніхто не звертає на твої слова увагу!	Чоловік	Християнин	40-49
Навіщо ти тут? Ти все одно нічого не зможеш зробити. Всі тільки сміються з тебе, а ти навіть не помічаєш цього.	Жінка	Буддист	50-59
Ми ж не в дитячому садку! Ти не вмієш нормально працювати, а постійно лише вносимеш хаос у все.	Чоловік	Єврей	10-19
Чого ти чекаєш? Тут всі знають, що ти не здатен на що-небудь серйозне, тому краще зупинись, поки не пізно.	Жінка	Мусульманин	20-29
Ти нічого не досягнеш. Таких як ти в житті не вистачає, щоб щось змінити.	Чоловік	Атеїст	30-39
Не розумію, чому ти вважаєш свої ідеї цінними. Часом здається, що ти просто втрачаєш час і всіх навколо!	Жінка	Християнин	40-49
Ти навіть не уявляєш, як це важко, тобі краще не вказувати, де і як потрібно працювати.	Чоловік	Буддист	50-59
У тебе нічого не вийде, якщо так продовжуватимеш. Це просто не твоя справа, не лізь!	Жінка	Єврей	10-19
Ти так і не зрозумів, що не все на світі дається так просто? Треба більше вчитися і працювати!	Чоловік	Християнин	20-29

Рис. 3.12 – Програмна реалізація перегляду міток текстових зразків датасету

У таблиці на рис. 3.12 наведено приклади текстових зразків з мітками, що відповідають етичним аспектам принципу справедливості FATE. Поля містять сам текстовий зразок, визначені мітки щодо етичних аспектів, такі як стать, релігійна приналежність і вікова група. Надання таких міток кожному тестовому зразку у датасеті дозволяє в подальшому оцінити відхилення фактичних пропорцій від цільових, що забезпечуватимуть репрезентативність датасету.

На рис. 3.13 наведено результат коригування датасету до цільових пропорцій згідно з популяцією України. Відповідно, було розв’язано багатокритеріальну задачу оптимізації та сформовано репрезентативний датасет згідно з етичними аспектами, такими як вік та гендер.



Розподіл Зразків

Розподіл зразків у сформованій репрезентативній вибірці після аугментації даних в результаті розв'язку багатокритеріальної оптимізаційної задачі:

Вікові демографічні підгрупи	0-19 років	20-29 років	30-39 років	40-49 років	50-100 років
Відсоткове відношення демографічних груп за гендером та віком у популяції України	9.67%	5.64%	8.96%	7.79%	15.56%
Чоловіки	9.67%	5.64%	8.96%	7.79%	15.56%
Жінки	9.04%	4.53%	7.96%	7.47%	23.38%
Відсоткове відношення демографічних груп за гендером та віком у текстовій вибірці	9.65%	5.62%	8.94%	7.80%	15.57%
Чоловіки	9.65%	5.62%	8.94%	7.80%	15.57%
Жінки	9.05%	4.57%	7.97%	7.45%	23.38%
Одержане відхилення від репрезентативного розподілу	0.02%	0.02%	0.02%	0.01%	0.02%
Чоловіки	0.02%	0.02%	0.02%	0.01%	0.02%
Жінки	0.01%	0.04%	0.01%	0.02%	0.00%

Зведення:

Отримано відхилення розподілів зразків за класами вікового та гендерного етичних аспектів датасету одночасно, трансформованого за створеним методом, від ідеального репрезентативного розподілу:

- Мінімальне: **0.00%**
- Максимальне: **0.04%**
- Середнє: **0.02%**

Рис. 3.13 – Результат коригування датасету до цільових пропорцій згідно з популяцією України

У результаті оцінювання та коригування репрезентативності досягнено цільових пропорцій зразків у датасеті з мінімальними відхиленнями. Далі такий датасет може використовуватись для навчання моделей, що виявлятимуть та класифікуватимуть типи кіберзалякувань. Візуальну аналітику щодо результатів коригування датасету до цільових пропорцій згідно з популяцією України наведено на рис. 3.14.

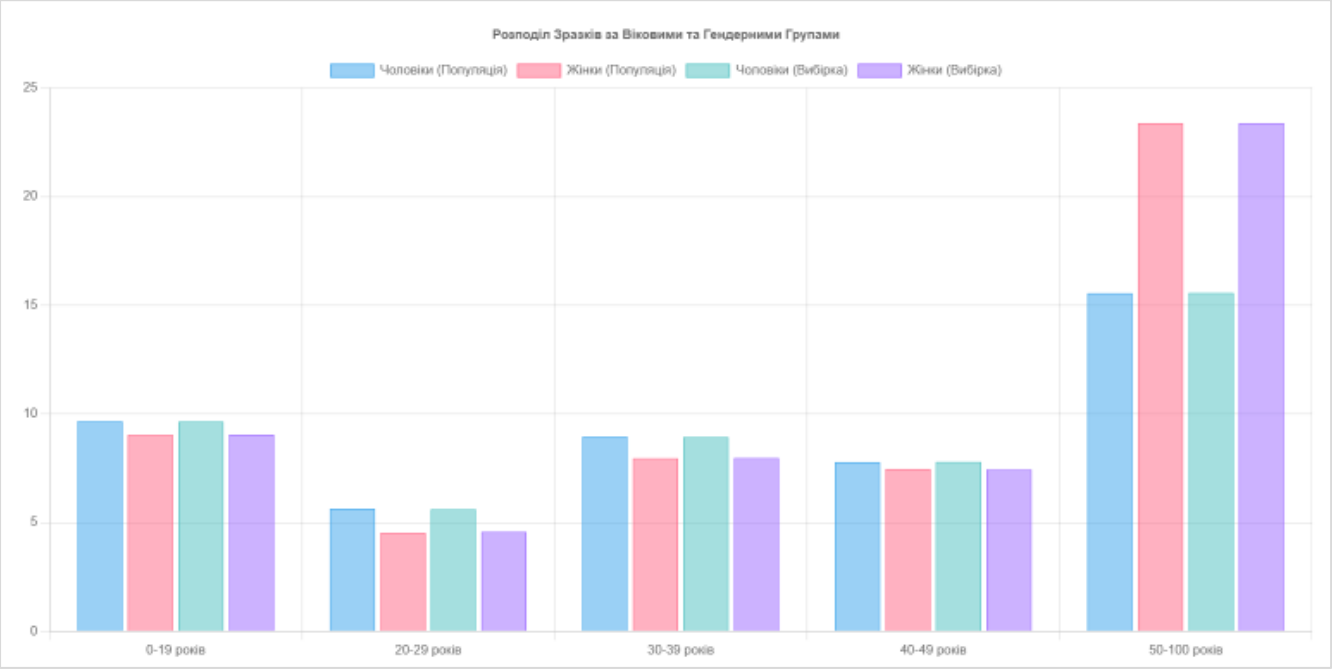


Рис. 3.14 – Візуальна аналітика щодо результатів коригування датасету до цільових пропорцій згідно з популяцією України

В рамках дослідження ефективності запропонованого методу виявлення і класифікації кіберзалякувань розроблена підсистема виявлення та класифікації кіберзалякувань у текстовому контенті (рис. 3.15) дозволяє визначити, чи має текстовий контент кіберзалякування і які типи кіберзалякувань присутні в досліджуваному текстовому контенті.

На рис. 3.15 подано графічний інтерфейс підсистеми виявлення та класифікації кіберзалякувань, де у верхній частині сторінки розташоване текстове поле з прикладом тексту для аналізу. Після натискання кнопки «Розпізнати» система здійснює класифікацію тексту за допомогою моделей глибокого навчання. У нижній частині інтерфейсу подано результати аналізу, які вказують на те, що текстовий контент «*Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести, тільки смішить жінки взагалі не повинні займатися програмуванням, особливо такі, як ти.*» визначений як такий, що містить в більшій мірі вікове та гендерне кіберзалякування. Проте, виведено також і

ймовірності наявності інших типів кіберзалякувань, які мають значно менші значення. Отримані результати демонструють, що модель здатна виявляти, а також класифікувати різні типи кіберзалякувань в одному текстовому зразку.

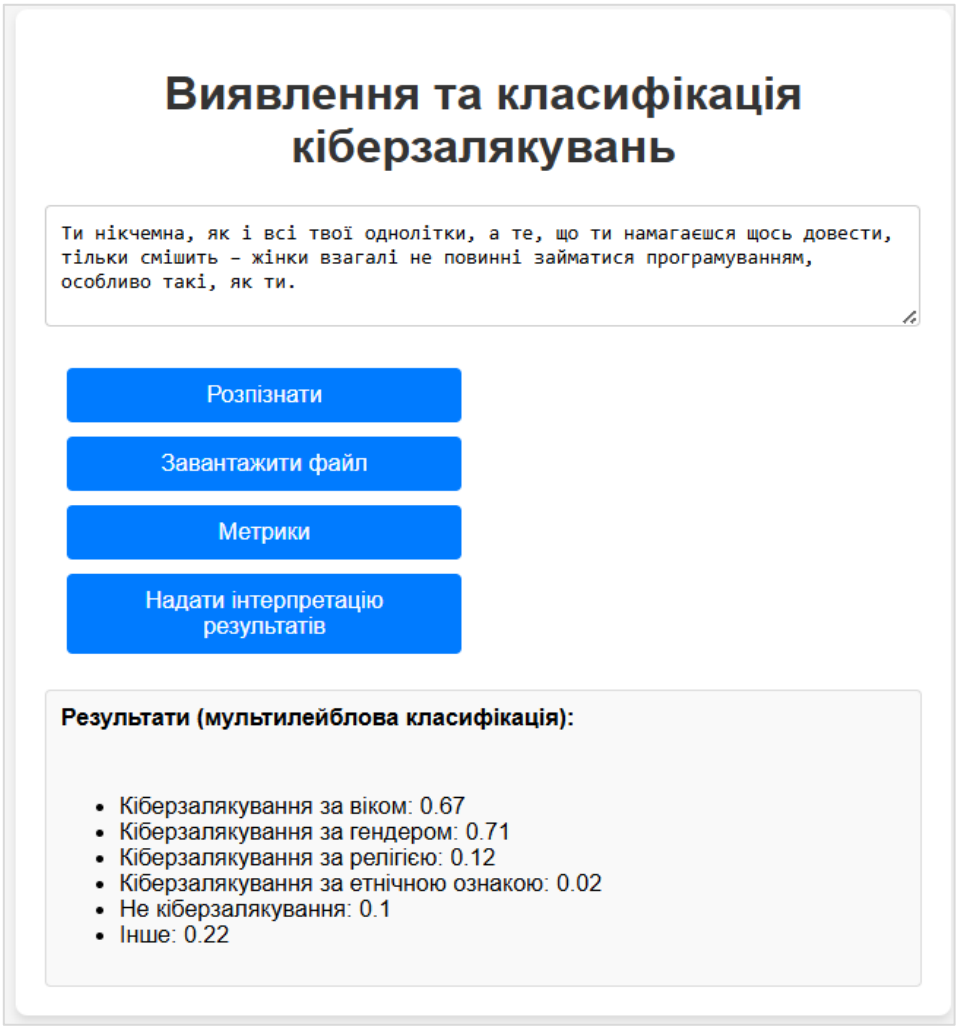


Рис. 3.15 – Результат виявлення кіберзалякувань у текстовому контенті

Після натиснення кнопки «Надати інтерпретацію результатів» відбувається інтерпретація результатів виявлених типів кіберзалякувань, яка реалізована у підсистемі інтерпретації результатів виявлення кіберзалякувань.

Для дослідження роботи підсистеми було використано текстовий контент «*Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести, тільки смішить – жінки взагалі не повинні займатися програмуванням, особливо*

такі, як ти». Було виявлено, що тестовий контент має ознаки кіберзалякувань, тому в ньому було класифіковано наступні типи кіберзалякувань:

– вікове: 67 %;

– етнічне: 2 %;

– гендерне: 71 %;

– інший тип: 22 %;

– релігійне: 12 %.

Шляхом застосування методу LIME для інтерпретації результатів моделі BERT, було отримано результати візуальної інтерпретації виявлених типів кіберзалякувань з використанням абсолютного значення ваг, що подані на рис. 3.16. Для інтерпретації прийнятих рішень моделлю BERT слова виділяються кольорами – найбільш яскравий колір означає найбільше значення ваги слова, тобто це слово мало найбільший вплив, найсвітліше – найменший.

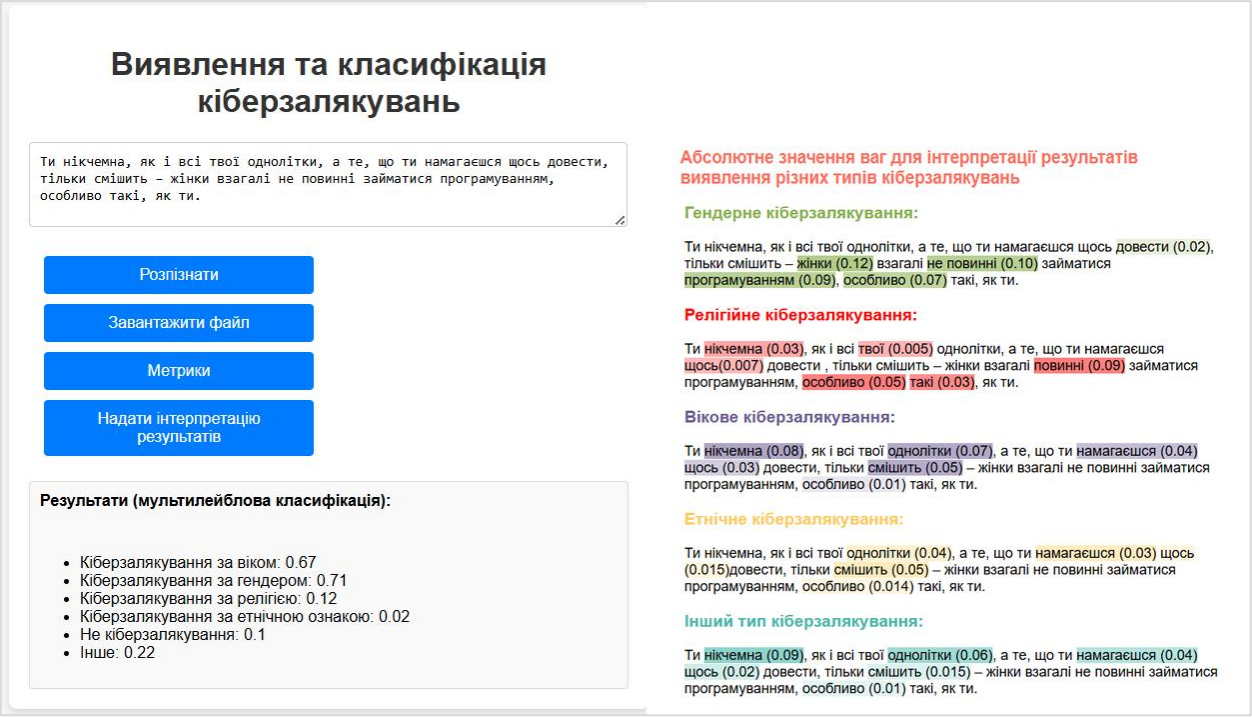


Рис. 3.16 – Використання абсолютного значення ваги для визначення яскравості кольору для інтерпретації результатів виявлення різних типів кіберзалякувань

Як показано на рис. 3.16, усі слова мають значення ваг є додатними. У цьому варіанті візуалізації яскравість кольору визначається абсолютним значенням ваги. Це призводить до однакового рівня виділення кольором як для слів, що мали позитивний вплив на прийняття рішення до віднесення текстового зразка до певно типу кіберзалякування, так і негативний вплив. Проте LIME надає як додатні, так і від'ємні ваги, які свідчать про зменшення ймовірності певного класу, однак обидва типи значень однаковою мірою впливають на остаточне рішення моделі. При цьому величина ваги, незалежно від її знака, вказує на ступінь впливу слова.

Так як для LIME важливо не лише демонструвати силу впливу слова, а й чітко розрізняти його характер: позитивний чи негативний, то з метою демонстрації цього характеру реалізовано альтернативне подання кольору та його яскравості, за яким від'ємні значення відображатимуться в іншій кольоровій гамі, ніж додатні. Такий спосіб представлення візуальної інтерпретації результатів ілюструється на рис. 3.17.

Використання окремих колірних схем для додатних і від'ємних значень у візуалізації інтерпретацій LIME є обґрунтованим з точки зору сприйняття даних та аналізу результатів моделей машинного навчання. Оскільки від'ємні ваги зменшують ймовірність належності до певного класу, а додатні її підвищують, застосування однакових кольорів для обох типів значень може спричинити некоректне трактування висновків кінцевими користувачами. Без додаткових засобів інтерпретації значення однакової інтенсивності, але протилежного знака, можуть виглядати рівнозначними, хоча їхній вплив на рішення моделі суттєво відрізняється.

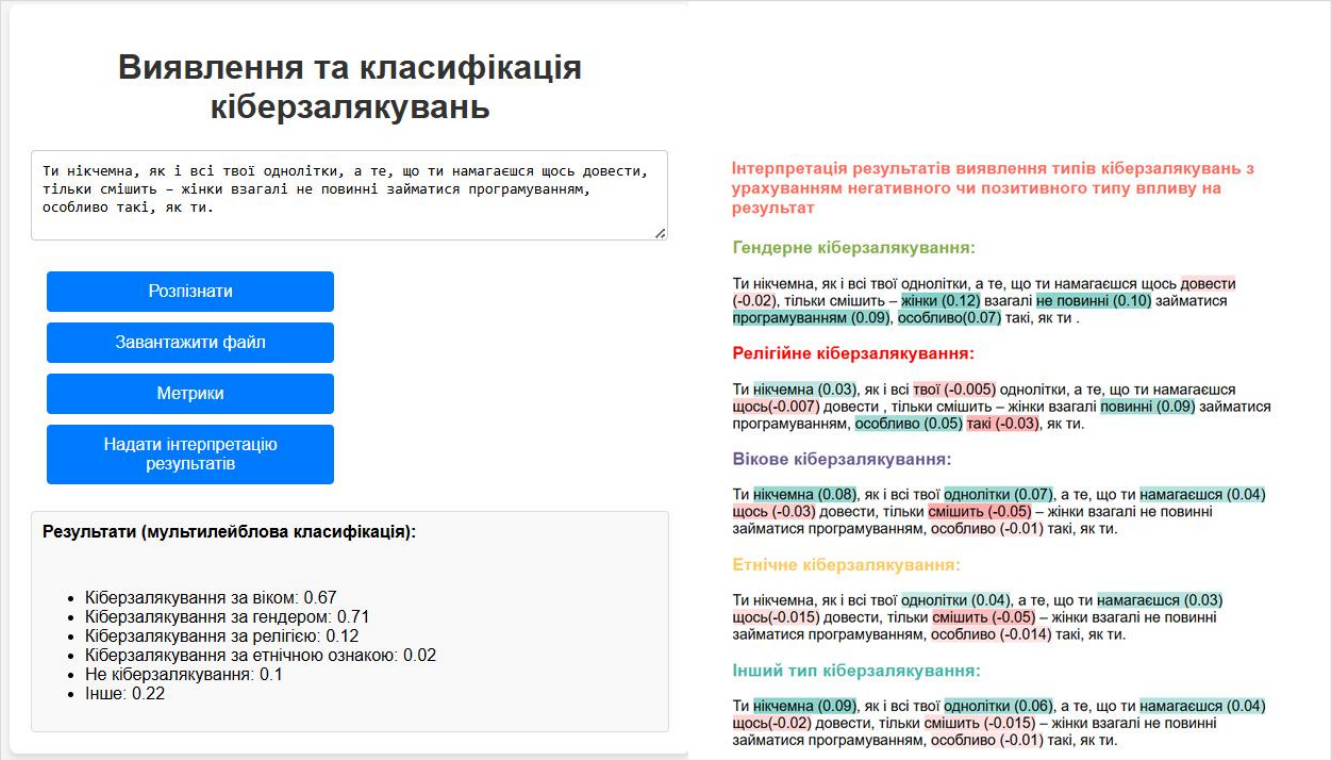


Рис. 3.17 – Визначення кольору та його яскравості
для інтерпретації результатів виявлення типів кіберзалякувань
з урахуванням негативного чи позитивного типу впливу на результат

Додатково створено діаграми для графічної інтерпретації впливу окремих слів текстового зразка на ймовірність віднесення його до конкретного типу кіберзалякування (рис. 3.18). Діаграми демонструють, як модель оцінює значущість кожного слова в тексті, враховуючи його внесок у результуюче рішення. Вплив слів відображається у вигляді горизонтальних стовпців, довжина яких відповідає величині впливу (ваги), а колір вказує на напрямок цього впливу. Коричневі стовпці позначають негативний вплив, тобто зменшення ймовірності віднесення тексту до певного класу, тоді як сині стовпці вказують на позитивний вплив, збільшуючи цю ймовірність. Величина впливу представлена числовими значеннями на горизонтальній осі діаграми.

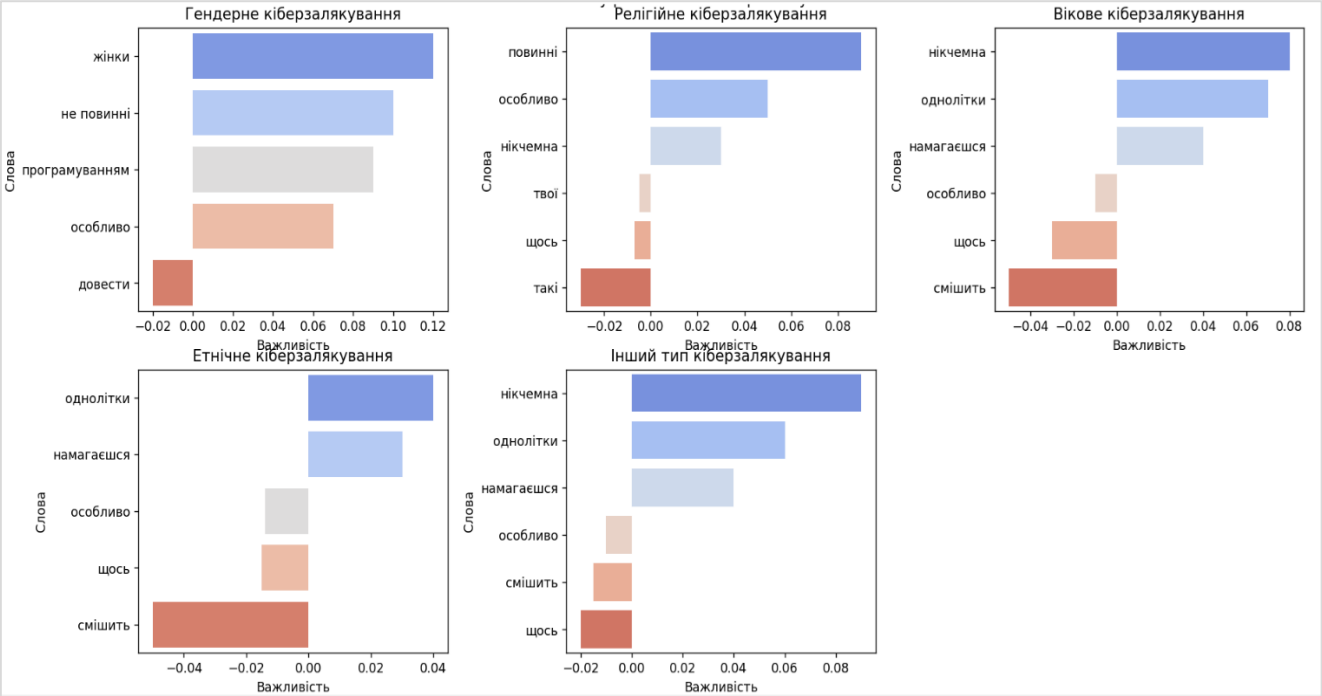


Рис. 3.18 – Діаграми для графічної інтерпретації впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування

Крім того у підсистемі визначається і середнє значення важливості кожного слова для всіх класів, що дозволяє оцінити його загальний вплив незалежно від конкретного типу кіберзалякування. Отримані значення візуалізовано у відповідній діаграмі (рис. 3.19).

Наприклад, терміни, що стосуються різних форм кіберзалякування, можуть мати високу вагу для кількох класів. Якщо слово має значний загальний вплив, це може свідчити про його універсальну роль у контексті кіберзалякування. Зокрема, слова, що позначають етнічну приналежність або релігію, можуть бути значущими для кількох класів, таких як «етнічне кіберзалякування» та «релігійне кіберзалякування», що вказує на потенційну крос-модальність ознак, які використовує модель. Водночас, якщо слово має високий вплив лише в одному класі, це підкреслює його специфічність і може свідчити про унікальні мовні патерни для цього типу кіберзалякування.

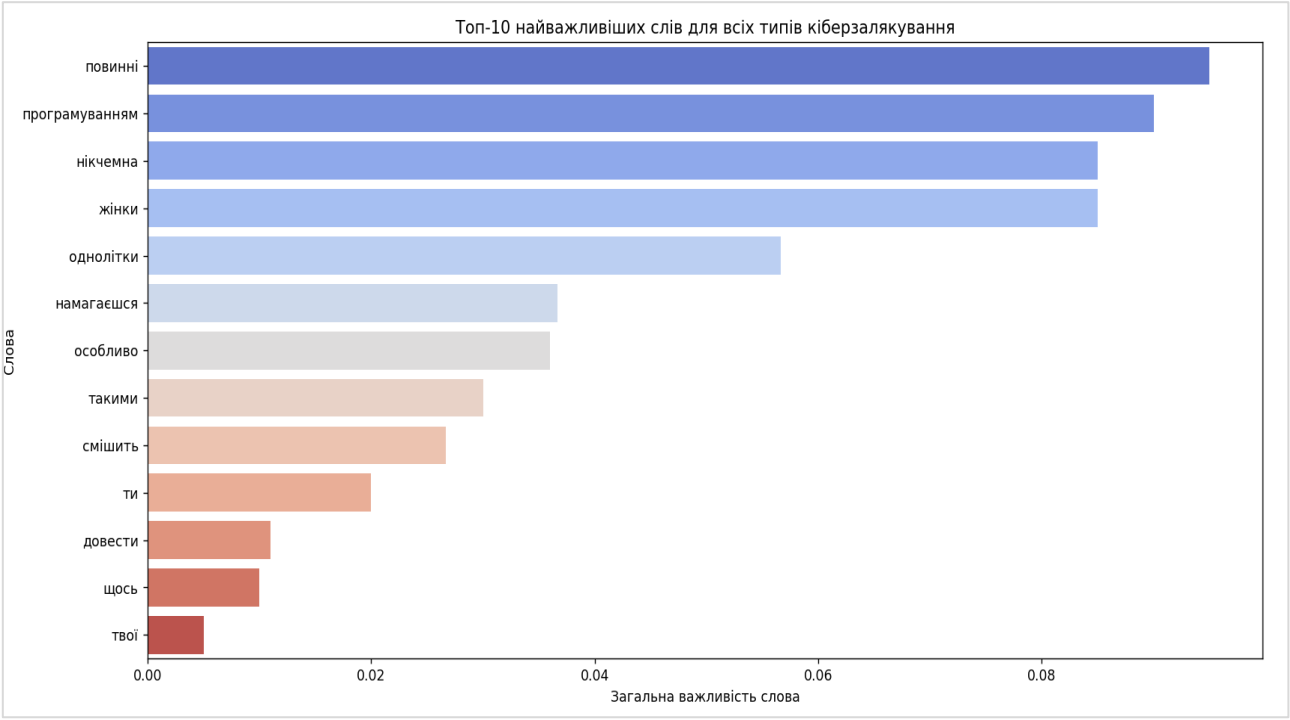


Рис. 3.19 – Діаграма з відображенням середнього значення важливості топ 10 слів для всіх класів

Наведеним чином було експериментально реалізовано візуальну інтерпретацію результатів виявлення та класифікації кіберзалякувань, які дозволяють оцінити, модель орієнтується на релевантні ознаки або її рішення були спричинені випадковими чи нерелевантними факторами. Наприклад, якщо в тексті виявляються слова, що не мають змістового зв’язку з віковим кіберзалякуванням, але вони отримують високу вагу, то це може свідчити про можливі помилки або набуті упередження в моделі.

3.7. Статистичні критерії оцінювання моделей класифікації

Статистичні показники, наведені далі, використовувались для оцінювання якості розроблених моделей машинного навчання для виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту.

Матриця помилок є однією з важливих оцінок ефективності моделей класифікації, що особливо актуально у випадку задачі виявлення та класифікації кіберзалякувань у текстовому контенті [133]. Вона дозволяє візуалізувати результативність моделі шляхом зіставлення реальних класів з передбаченими, що сприяє отриманню докладного аналізу її роботи. Кожен елемент матриці представляє кількість випадків, що належать до певної комбінації передбачень та реальних міток: правильні позитивні (True Positive), правильні негативні (True Negative), хибні позитивні (False Positive) та хибні негативні (False Negative) [134].

Застосування матриці помилок для оцінки моделей виявлення та класифікації кіберзалякувань у текстовому контенті дозволяє проаналізувати ефективність моделі на рівні кожного окремого класу. Оскільки текстові дані з кіберзалякувань часто характеризуються нерівномірним розподілом між категоріями (наприклад, класи, пов'язані з віковими чи етнічними образами, можуть бути представлені неоднаково), матриця помилок надає змогу оцінити, наскільки модель справляється з такими дисбалансами [113].

Матриця помилок також є основою для розрахунку важливих статистичних показників, таких як точність (Accuracy), влучність (Precision), повнота (Recall) та F1-міра [136].

Accuracy – точність, частка правильно класифікованих прикладів від загальної кількості:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.1}$$

де TP – кількість правильно передбачених позитивних випадків; TN – кількість правильно передбачених негативних випадків; FP – кількість хибно передбачених позитивних випадків; FN – кількість хибно передбачених негативних випадків. Діапазон: [0, 1], де 1 означає ідеальну точність, а 0 – повна відсутність правильних передбачень.

Precision – влучність, частка правильних позитивних передбачень серед усіх передбачених позитивних випадків:

$$Precision = \frac{TP}{TP + FP}, \tag{3.2}$$

Діапазон: [0, 1], де 1 означає ідеальну повноту (усі позитивні передбачення правильні).

Recall – повнота, частка правильних позитивних передбачень серед усіх реальних позитивних випадків:

$$\text{Re}ccal = \frac{TP}{TP + FN}. \tag{3.3}$$

Діапазон: [0, 1], де 1 означає ідеальну чутливість (всі справжні позитивні випадки виявлені).

F_1-score – середнє гармонійне між Precision та Recall, обчислюється за формулою:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{3.4}$$

Діапазон: [0, 1], де 1 вказує на ідеальний баланс між влучністю та повнотою [136].

Отже, для оцінки якості розроблених моделей для виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту використано такі статистичні критерії оцінювання матриця помилок, Accuracy, Precision, Recall, F1-score.

3.8. Висновки до розділу 3

Реалізована інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань, побудована на основі розроблених методів,

дозволяє експериментально оцінити якість та точність виявлення кіберзалякувань у текстовому контенті за допомогою засобів штучного інтелекту.

Спроєктовано архітектуру інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань, яка містить чотири підсистеми: оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості (Dataset Evaluation and Adjustment Subsystem), виявлення кіберзалякувань у текстовому контенті (Cyberbullying Detection Subsystem), інтерпретації результатів виявлення класифікації кіберзалякувань (Explanation and Visualization Subsystem), навчання моделей (Model Training Subsystem).

На етапі оцінки репрезентативності датасету для різних етичних аспектів було обрано відповідні нейромережеві моделі: LSTM – для гендерного аспекту, BERT – для релігійного, а SVM – для вікового. Для виявлення та класифікації кіберзалякувань використано дві нейромережеві моделі: архітектура BiLSTM застосовано для бінарної класифікації, тоді як архітектуру BERT використано для мультилейблової класифікації типів кіберзалякувань.

Таким чином, реалізована інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань дозволяє здійснювати виявлення та класифікацію кіберзалякувань у текстовому контенті, забезпечуючи репрезентативність навчальних датасетів та інтерпретацію отриманих результатів. Розроблена інформаційна система призначена для експериментального дослідження розроблених методів виявлення та класифікації кіберзалякувань у текстовому контенті, яке наведено у розділі 4.

РОЗДІЛ 4.
ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ

У розділі описано експериментальні дослідження розроблених методів виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту. Для цього використано наведені у пункті 4.1 статистичні критерії оцінювання. Досліджено нейромережеві архітектури, що використовуються для аналізу текстового контенту. Отримані результати порівняно з результатами відомих досліджень.

Для дослідження методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості навчено моделі машинного і глибокого навчання та порівняно отримані статистичні показники. Також наведено результати аналізу на репрезентативність датасету для виявлення кіберзалякувань, відповідно до яких здійснено його коригування згідно з вимогами.

Для дослідження методу виявлення кіберзалякувань у текстовому контенті навчено моделі глибокого навчання як для бінарної класифікації, так і для мультилейблової, й порівняно отримані результати.

З метою дослідження методу інтерпретації результатів виявлення кіберзалякувань, наведено приклади аналізу текстових зразків, без урахування обмежень, що описані у пункті 2.2 для виявлення характеру впливу цих обмежень.

4.1. Експериментальне дослідження методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

Для дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості навчено декілька моделей машинного навчання. Результати обчислення статистичних метрик як

Accuracy, Precision, Recall та F1-score моделей для гендерного етичного аспекту наведено в таблиці 4.1, результат фактичного балансу класів текстової вибірки кіберзалякувань за гендерним етичним аспектом наведено на рис. 4.1.

Таблиця 4.1

Статистичні метрики Accuracy, Precision, Recall та F1-score моделей машинного навчання за гендерним етичним аспектом

Модель	Accuracy, %	Precision, %	Recall, %	F1-score, %
Гендерний етичний аспект				
FastForest	55	56	52	54
SVM	50	50	50	50
LSTM	75	79	73	76
BERT	59	55	60	57

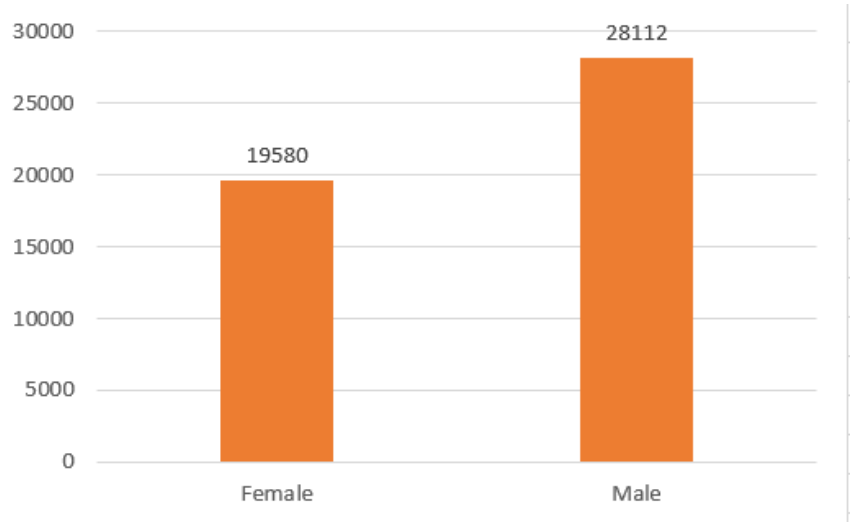


Рис. 4.1 – Баланс розподілу вхідного датасету за гендерним етичним аспектом FATE-принципу справедливості

Результати обчислення статистичних метрик моделей машинного навчання для вікового етичного аспекту наведено в таблиці 4.2, результат фактичного балансу класів текстової вибірки кіберзалякувань за віковим етичним аспектом наведено на рис. 4.2.

Таблиця 4.2

Статистичні метрики Accuracy, Precision, Recall та F1-score моделей машинного навчання за віковим етичним аспектом

Модель	Accuracy, %	Precision, %	Recall, %	F1-score, %
Віковий етичний аспект				
FastForest	46	46	43	43
SVM	79	68	59	63
LSTM	51	52	48	50
BERT	50	37	39	38

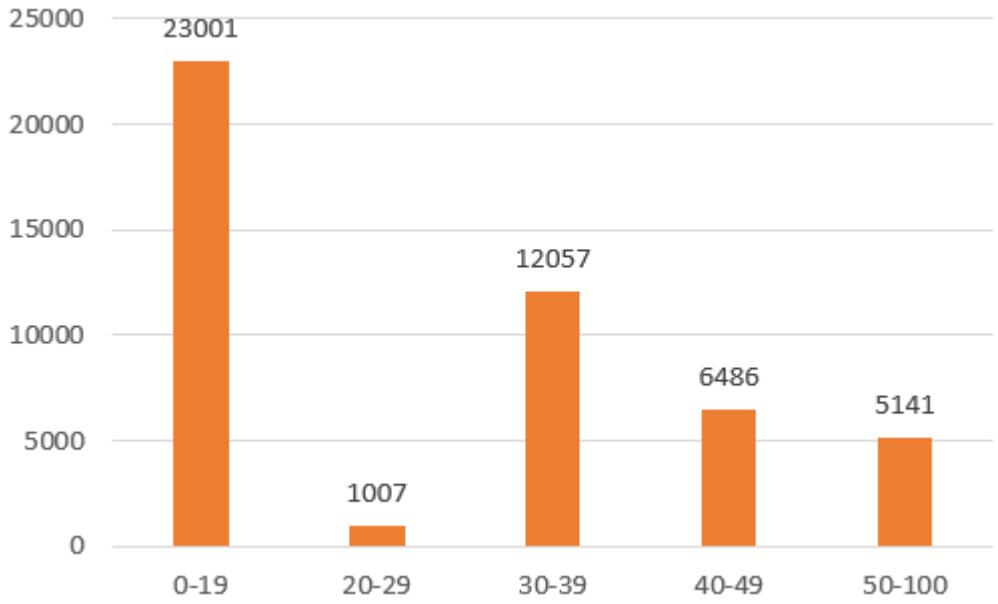


Рис. 4.2 – Баланс розподілу вхідного датасету за віковим етичним аспектом FATE-принципу справедливості

Результати обчислення статичних метрик моделей машинного навчання для релігійного етичного аспекту наведено в таблиці 4.3, результат фактичного балансу класів текстової вибірки кіберзалякувань за релігійним етичним аспектом – на рис. 4.3.

Таблиця 4.3

Статистичні метрики Accuracy, Precision, Recall та F1-score
моделей машинного навчання за релігійним етичним аспектом

Модель	Accuracy, %	Precision, %	Recall, %	F1-score, %
Релігійний етичний аспект				
FastForest	56	58	55	56
SVM	60	62	59	60
LSTM	61	68	61	62
BERT	85	87	71	78

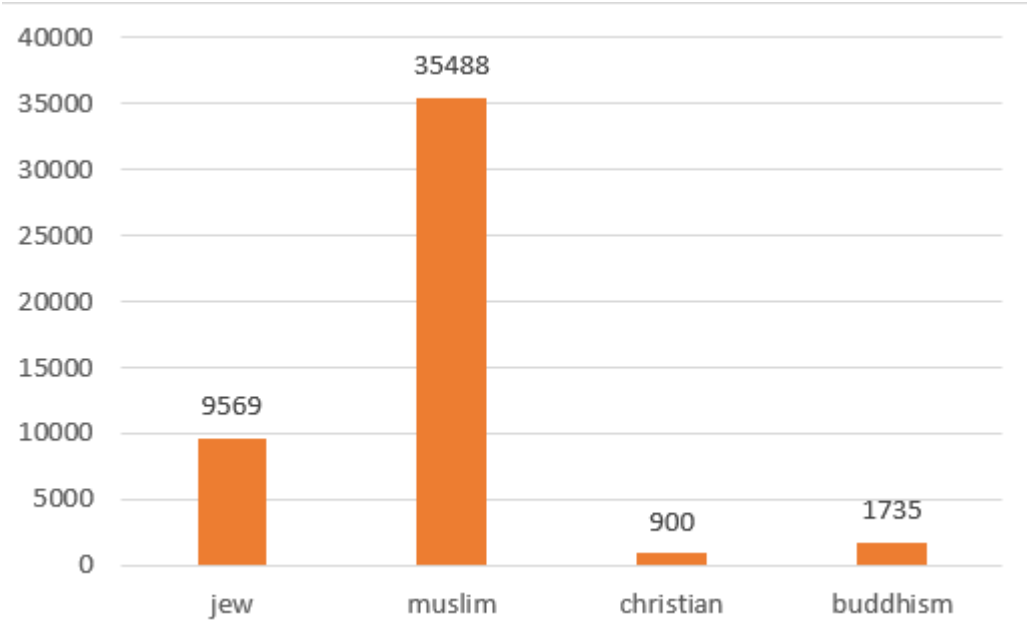


Рис. 4.3 – Баланс розподілу вхідного датасету
за релігійним етичним аспектом FATE-принципу справедливості

Слід відмітити, що для різних класів було отримано різні рівні лінійної роздільної здатності: за релігійним аспектом з використанням класифікатора BERT, який показав найкращий результат з навчених моделей, дані виявились добре роздільні, за гендерним аспектом з використанням класифікатора LSTM

дані виявились середньороздільні та за віковим аспектом з використанням класифікатора SVM – погано роздільні.

Окрім того, згідно з діаграмами результатів (рис. 4.1–4.3), що демонструють баланс вхідного датасету за трьома етичними аспектами, виявлено що датасет не є репрезентативним, адже класи різних етичних аспектів мають кількість текстових зразків, що не відповідає пропорціям демографічних підгруп населення України, які були взяті як цільові, таким чином потребують збалансування для набуття репрезентативного вигляду. Тому, згідно з кроками методу метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, датасет потребує аугментації даних. Для цього необхідно вирішити оптимізаційну задачу, для коректного видалення надлишкових елементів кожного класу за кожним з етичних аспектів з подальшою аугментацією вибірки даних до цільових вимог (кількість елементів та пропорції класів).

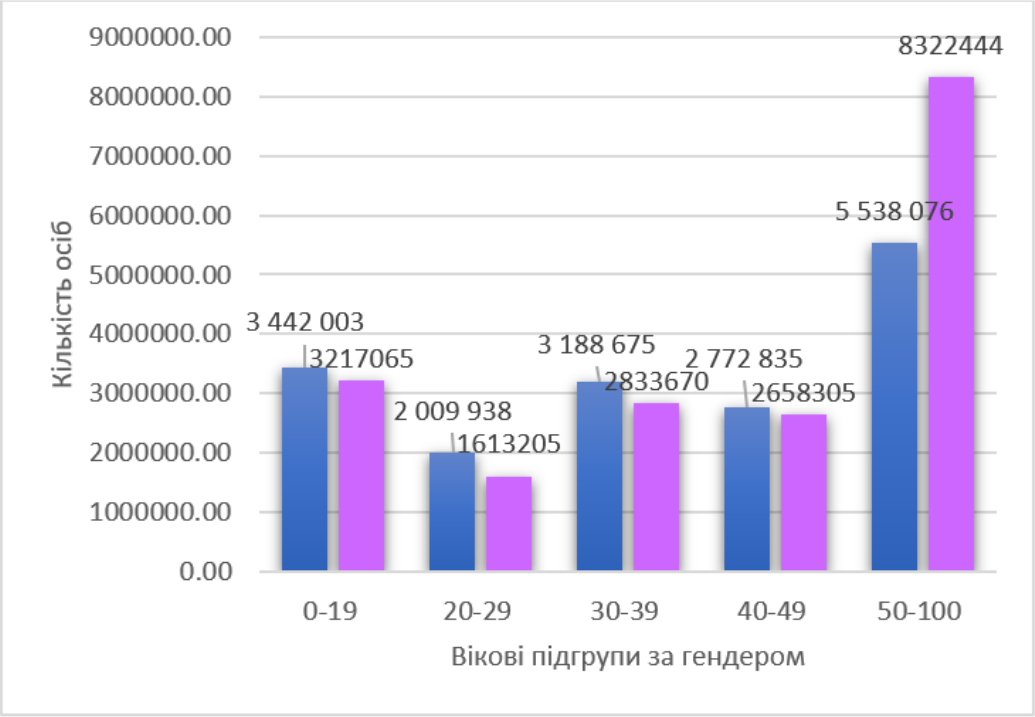


Рис. 4.4 – Статистика віку жінок та чоловіків в Україні на 2023 рік за [144]

Для аналізу та формування репрезентативного датасету за цільові пропорції класів для формування репрезентативної вибірки текстових даних за віком та гендером взято популяцію України. За даними Інституту демографії та соціальних досліджень імені М. В. Птухи Національної академії наук України [143] станом на липень 2023 року загальна чисельність населення України оцінюється в 35596216 осіб. У кожній віковій підгрупі представлено наступну кількість осіб: вікова група 0–19 років – 6 659 068 осіб, 20–29 років – 3 623 143 осіб, 30–39 років – 6 022 345 осіб, 40–49 років – 5 431 140 осіб, 5–100 років – 13 860 520 осіб. Щодо гендерної структури населення України на 2023 рік: 16951527 – жінки, а 18644689 – чоловіки [144], детальніша статистика наведена на рис. 4.4. Варто зауважити, що в межах цієї роботи при аналізі гендерного етичного аспекту розглядається цисгендерна група.

Таблиця 4.4

Відсоткові відношення зразків за віком у датасеті та осіб популяції у вікових демографічних підгрупах

Вікові демографічні підгрупи	**Відсоток зразків за віком у датасеті, %**	**Відсоток осіб популяції у вікових демографічних підгрупах, %**	**Відхилення текстових зразків від підгруп популяції, %**	**Новий розподіл класів датасету, %**	**Відхилення від репрезен-тативного розподілу, %**
0–19 років	48,23	18,71	29,52	18,75	**0,04**
20–29 років	2,11	10,17	8,06	10,15	**0,02**
30–39 років	25,28	16,92	8,36	16,87	**0,03**
40–49 років	13,60	15,26	1,66	15,28	**0,02**
50–100 років	10,78	38,94	28,16	38,95	**0,01**

У таблиці 4.4 подано відсоткові відношення зразків за віком у датасеті та осіб популяції у вікових демографічних підгрупах, а також обчислено новий

розподіл класів датасету, якби враховувався лише один етичний аспект – віковий. У таблиці 4.5 наведено відсоткові відношення зразків за гендером у датасеті та осіб популяції у гендерних демографічних підгрупах, а також обчислено новий розподіл класів вибірки, якби враховувався лише один етичний аспект – гендерний.

Таблиця 4.5

Відсоткові відношення зразків за гендером у датасеті та осіб популяції у гендерних демографічних підгрупах

Гендерні демо-графічні підгрупи	**Відсоток зразків за гендером у датасеті, %**	**Відсоток осіб популяції у гендерних демографічних підгрупах, %**	**Відхилення текстових зразків від підгруп популяції, %**	**Новий розподіл класів датасету, %**	**Відхилення від репрезен-тативного розподілу, %**
Чоловіки	58,94	43,28	15,67	43,25	**0,03**
Жінки	41,06	56,72	15,67	56,75	**0,03**

Одержано відхилення розподілів зразків за класами вікового етичного аспекту датасету, трансформованого за створеним методом, від ідеального репрезентативного розподілу склали: мінімальне 0,01 %, максимальне 0,04 %, середнє 0,02 %, а для гендерного етичного аспекту: мінімальне 0,03 %, максимальне 0,03 %, середнє 0,03 %.

Проте, оптимізаційна задача з формування репрезентативної датасету є багатокритеріальна, критеріями в якій є формування датасету за віковим та гендерним етичним аспектом, тому метою є мінімізація відхилення між поточними та бажаними співвідношеннями класів, враховуючи обмеження на кількість зразків і можливості генерації нових даних.

В результаті вирішення оптимізаційної задачі для формування репрезентативного датасету за віковим та гендерним етичними аспектами на прикладі демографічних підгруп популяції України, отримано шляхом аугментації репрезентативний датасет, розподіл зразків у якого подано у таблиці 4.6, відповідний баланс класів датасету відображено на рис. 4.5 та 4.6.

Таблиця 4.6

Розподіл зразків у сформованому репрезентативному датасеті після аугментації даних в результаті розв'язку багатокритеріальної оптимізаційної задачі

Вікові демографічні підгрупи	**0–19 років**	**20–29 років**	**30–39 років**	**40–49 років**	**50–100 років**
Відсоткове відношення демографічних груп за гендером та віком у популяції України					
Чоловіки	9,67 %	5,64 %	8,96 %	7,79 %	15,56 %
Жінки	9,04 %	4,53 %	7,96 %	7,47 %	23,38 %
Відсоткове відношення демографічних груп за гендером та віком у датасеті					
Чоловіки	9,65 %	5,62 %	8,94 %	7,80 %	15,57 %
Жінки	9,05 %	4,57 %	7,97 %	7,45 %	23,38 %
Одержане відхилення від репрезентативного розподілу					
Чоловіки	**0,02 %**	**0,02 %**	**0,02 %**	**0,01 %**	**0,02 %**
Жінки	**0,01 %**	**0,04 %**	**0,01 %**	**0,02 %**	**0,00 %**

Отримано відхилення розподілів зразків за класами вікового та гендерного етичних аспектів датасету одночасно, трансформованого за запропонованим методом, від ідеального репрезентативного розподілу склали: мінімальне – 0,00 %, максимальне – 0,04 %, середнє – 0,02 %.

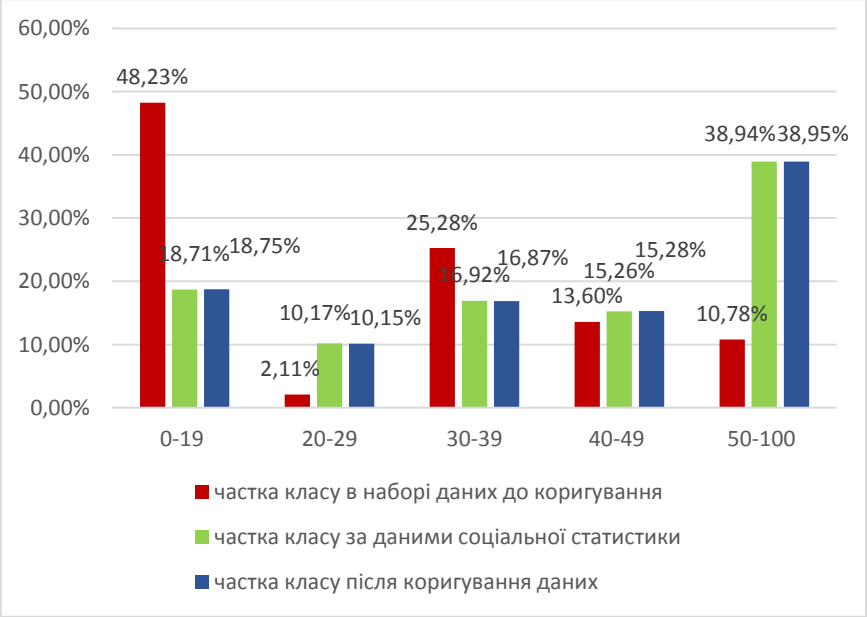


Рис. 4.5 – Баланс розподілу вхідного датасету
за віковим етичним аспектом FATE-принципу справедливості

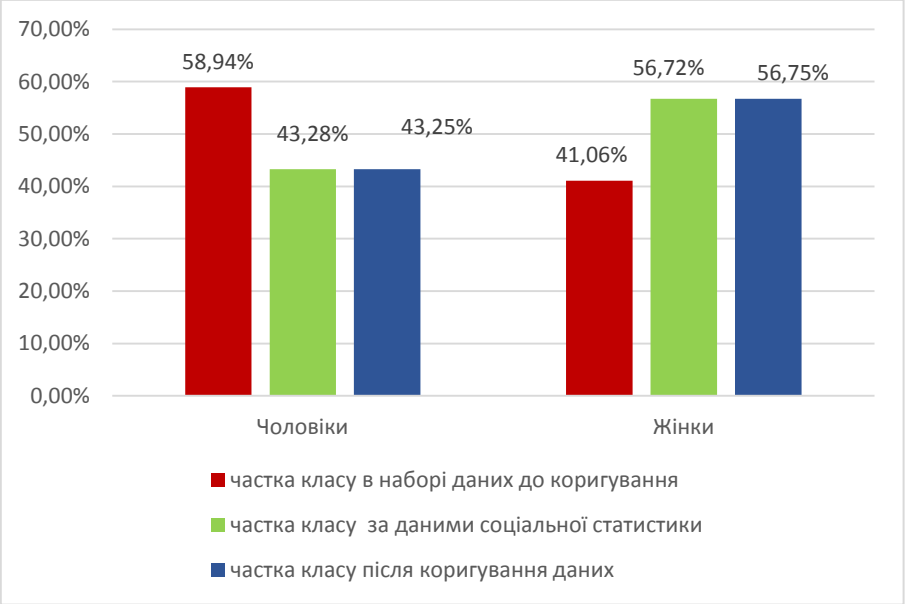


Рис. 4.6 – Баланс розподілу вхідного датасету
за гендерним етичним аспектом FATE-принципу справедливості

Отже, аналіз результатів класифікації моделей для етичних аспектів показав різний рівень точності. Найкращі показники точності було досягнуто для

релігійного етичного аспекту за допомогою моделі BERT, яка виявилася найбільш точною серед випробуваних підходів для класифікації текстів за релігійним етичним аспектом. Для гендерної ознаки класифікатор LSTM продемонстрував середні показники точності, перевершуючи інші моделі для вказаного етичного аспекту. Водночас, за віковим етичним аспектом модель SVM показала гірші показники точності, що свідчить про складність розділення вікових груп у текстовому контенті.

Також в результаті виконання кроків методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості сформовано датасет для виявлення кіберзалякувань, який є недискримінаційним та відображає пропорційне до реальних демографічних підгруп популяції України представлення зразків в датасеті.

4.2. Експериментальне дослідження методу виявлення кіберзалякувань у текстовому контенті

Для виявлення кіберзалякувань у текстовому контенті досліджено дві моделі глибокого навчання: BiLSTM та RNN з шаром LSTM, для цього реалізовано два різні підходи – проста бінарна класифікація за допомогою моделі BiLSTM та класифікація на основі словника та моделі RNN з шаром LSTM, який використовує оцінки зі словника як додаткові вхідні дані для LSTM, що підвищує точність виявлення кіберзалякувань за рахунок вловлювання складніших лексичних і контекстуальних зв'язків.

Для реалізації обох підходів використано мову програмування Python та бібліотеку Tensorflow. Результати метрик Accuracy, Precision, Recall та F_1, та наведено у таблиці 4.7.

Одержані результати статистичних метрик бінарної класифікації з використанням обох підходів перевершують результати в інших дослідженнях,

наприклад, у [83], у якому автори досягли точності 88 % з використанням моделі BERT, а також [146], у якому автори досягли найкращого показника точності 81 % за допомогою моделі Hate-BERT та у [147], автори якого досягли точності 93,2 % з використанням багатошарового перцептрона.

Таблиця 4.7

Підходи до бінарної класифікації

Модель	**Підхід до класифікації**	**Результати дослідження ефективності, %**
BiLSTM	Два класи: кіберзалякування і не кіберзалякування	Accuracy: 94 Precision: 94 Recall: 94 F_1-score: 94
RNN з LSTM шаром	Словник + нейронна мережа для аналізу сентименту	Accuracy: 94 Precision: 89,2 Recall: 91,2 F_1-score: 91,2

Для класифікації типів кіберзалякувань у текстовому контенті з використанням мультилейблової класифікації було навчено декілька нейромережевих моделей: BiLSTM, LSTM, GRU, RoBERTa, BERT. Для дослідження ефективності навчених моделей обраховано макро- та мікрометрики Acuracy, Precision, Recall, F_1. Обчислення як макро-, так і мікрометрик є важливим для комплексної оцінки моделей глибокого навчання, особливо в мультилейбловій класифікації для виявлення різних типів кіберзалякування в одному зразку.

Мікрометрики оцінюють точність моделі для кожного класу окремо, що особливо важливо в умовах дисбалансу класів. Макрометрики, з іншого боку, забезпечують загальну картину точності моделі, об’єднуючи інформацію з усіх класів. Комбінування обох типів метрик дозволяє отримати більш детальне

розуміння точності навченої моделі [148]. На рис. 4.7 подано діаграму з показниками макрометрик моделей.

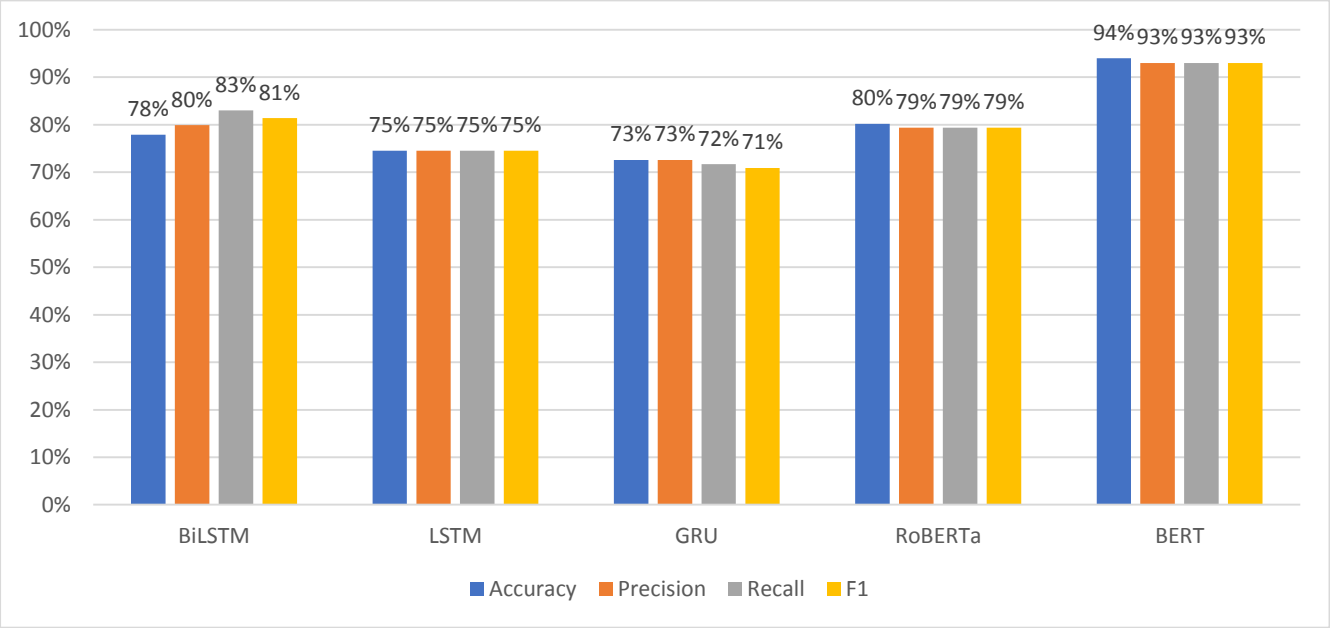


Рис. 4.7 – Показники макрометрик нейромережевих моделей

Відповідно до рис. 4.7, кращі серед інших результати отримала модель BERT, яка має показник точності 94 %, отже модель краще класифікує загальну кількість зразків без помилок. Також всі основні макрометрики BERT знаходяться на однаковому рівні – 93 %. Така стабільність означає, що модель добре збалансована в обох аспектах: і точності передбачення позитивних випадків, і здатності правильно виявляти всі наявні позитивні приклади. Так стабільність є важливою для завдання виявлення кіберзалякувань, де потрібно уникати як помилкових передбачень, так і упущених випадків кіберзалякувань.

Далі на рис. 4.8 – 4.11 показано діаграми з показниками мікрометрик моделей нейромережевих моделей для мультилейблової класифікації таких типів кіберзалякувань як віковий, релігійний, за етнічною приналежність та гендерний.

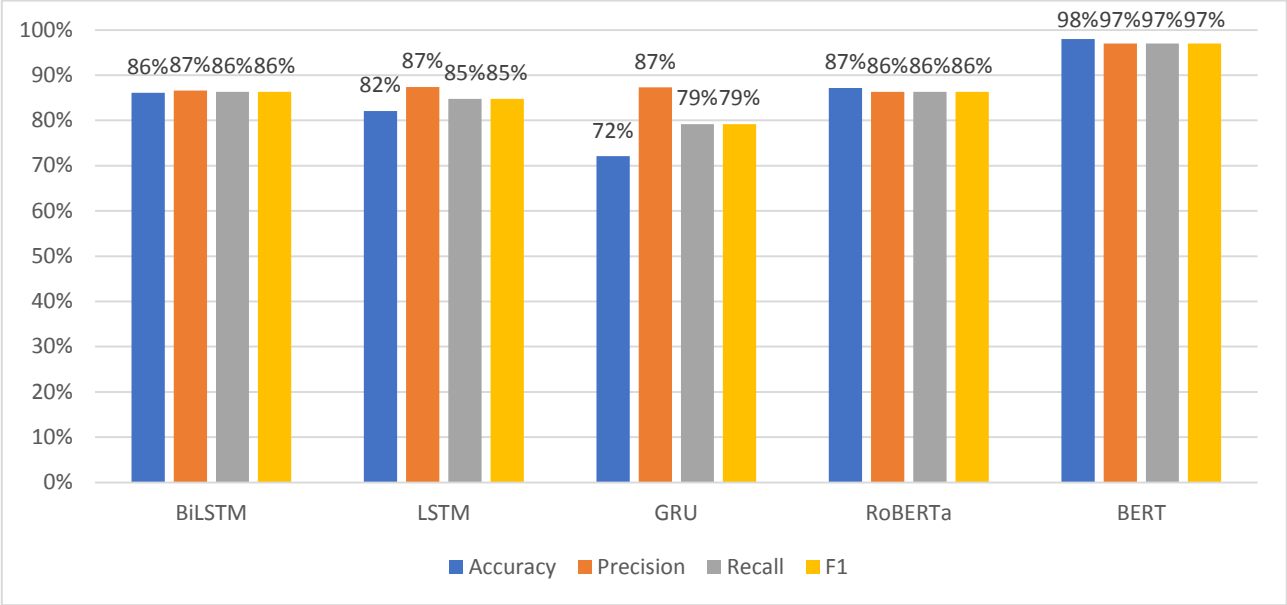


Рис. 4.8 – Показники мікрометрик нейромережевих моделей для мультилейблової класифікації типу кіберзалякування за віком

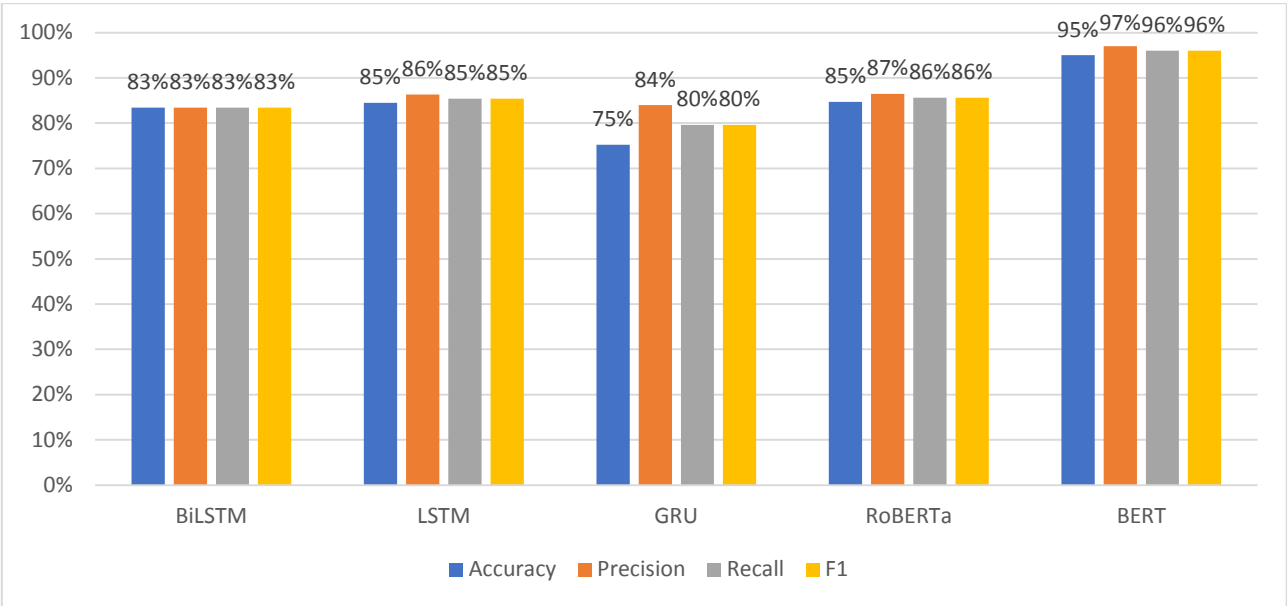


Рис. 4.9 – Показники мікрометрик нейромережевих моделей для мультилейблової класифікації типу кіберзалякування за релігією

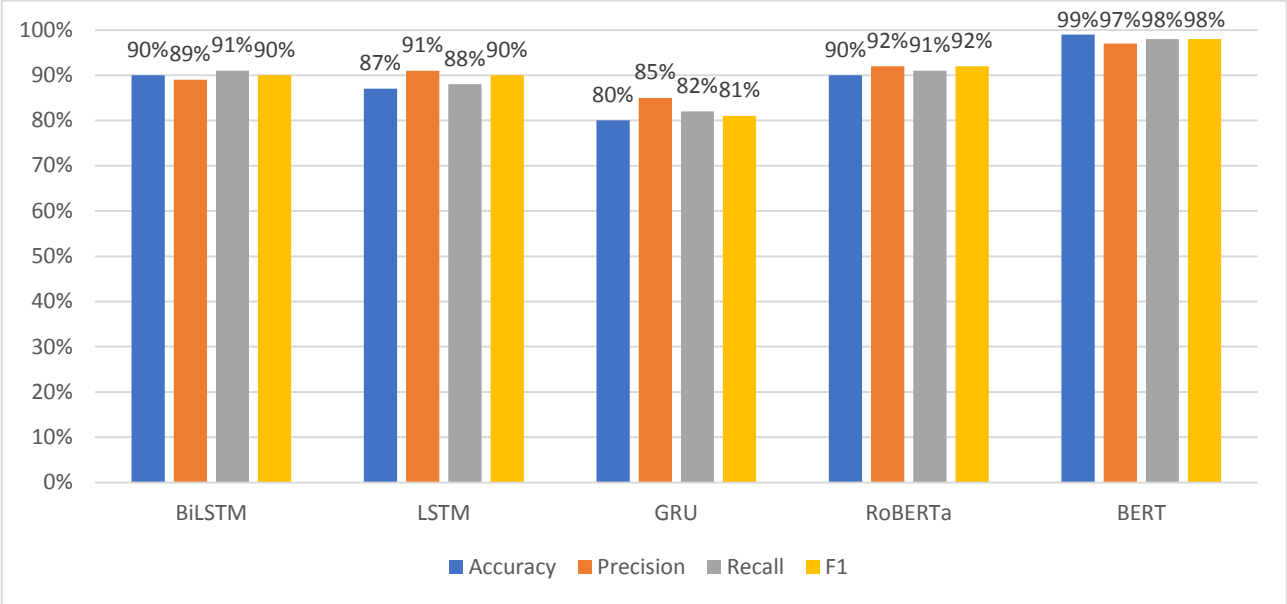


Рис. 4.10 – Показники мікрометрик нейромережевих моделей для мультилейблової класифікації типу кіберзалякування за етнічною приналежністю

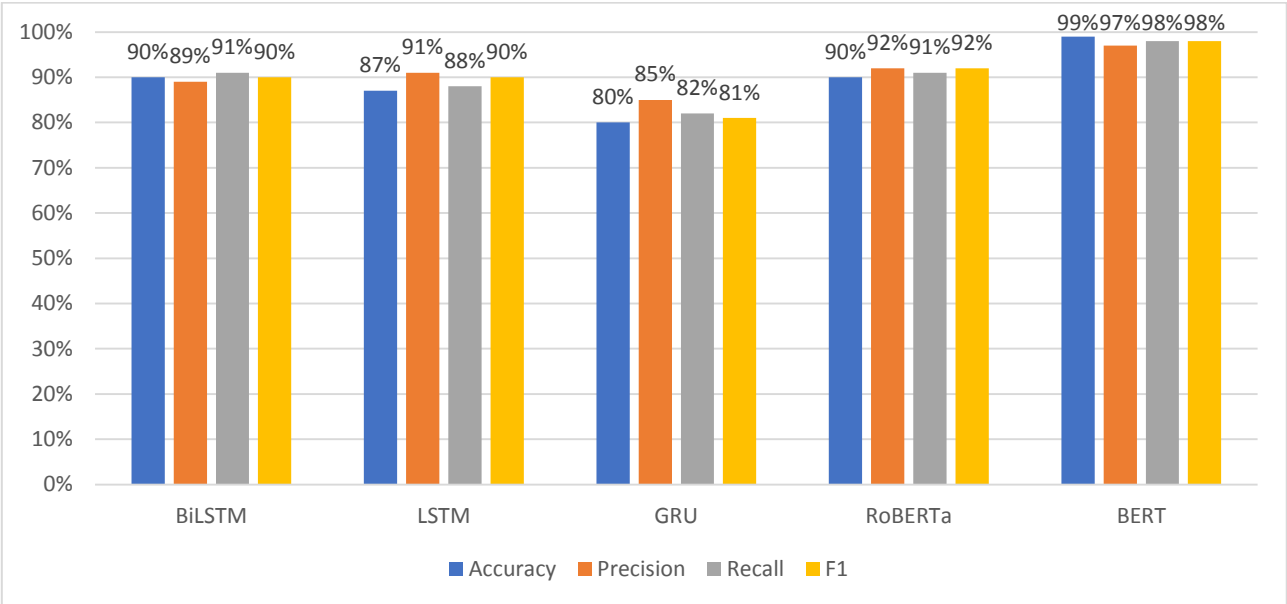


Рис. 4.11 – Показники мікрометрик нейромережевих моделей для мультилейблової класифікації типу кіберзалякування за гендером

Відповідно до наведених рис. 4.7 – 4.11, можна зробити висновок про те, що нейромережева модель BERT отримала кращі показники мікрометрик для виявлення усіх розглянутих типів кіберзалякувань, ніж інші нейромережеві моделі, що були досліджені. Також модель отримала кращі показники точності, ніж в існуючих підходах, наприклад, у [149] та [150], показники точності яких не перевищують 84 % та 88,3 %, відповідно.

Далі більш детально розглянуто статистику моделі BERT щодо правильних та неправильних прогнозів стосовно наведених у дослідженні типів кіберзалякування. На рис. 4.12 – 4.15 наведено розподіл коректно та некоректно класифікованих текстових зразків щодо різних типів кіберзалякування.

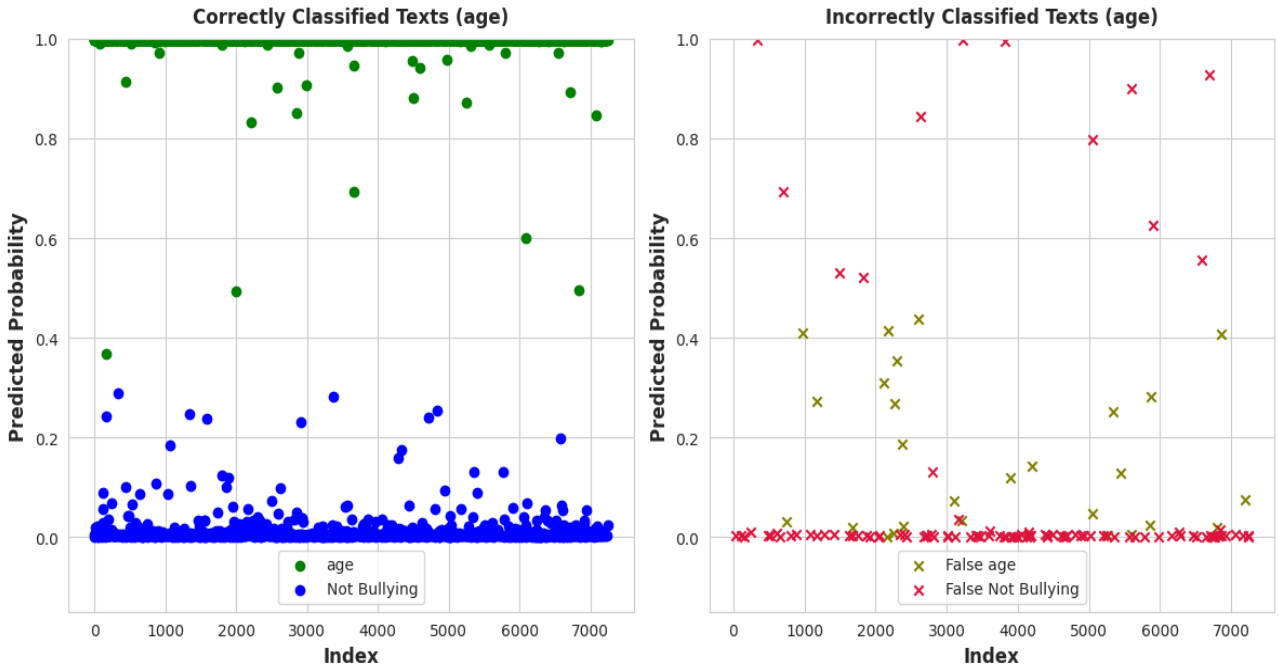


Рис. 4.12 – Розподіл коректно та некоректно класифікованих моделлю BERT текстових зразків щодо типу кіберзалякувань за віком

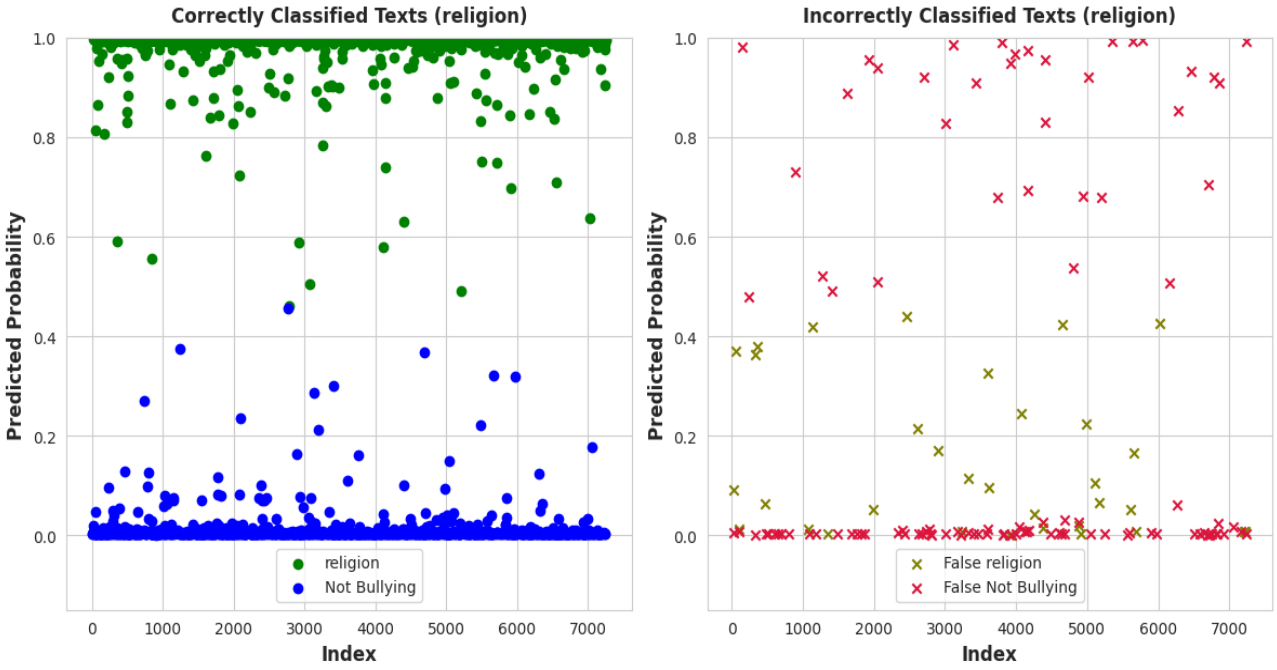


Рис. 4.13 – Розподіл коректно та некоректно класифікованих моделлю BERT текстових зразків щодо типу кіберзалякувань за релігією

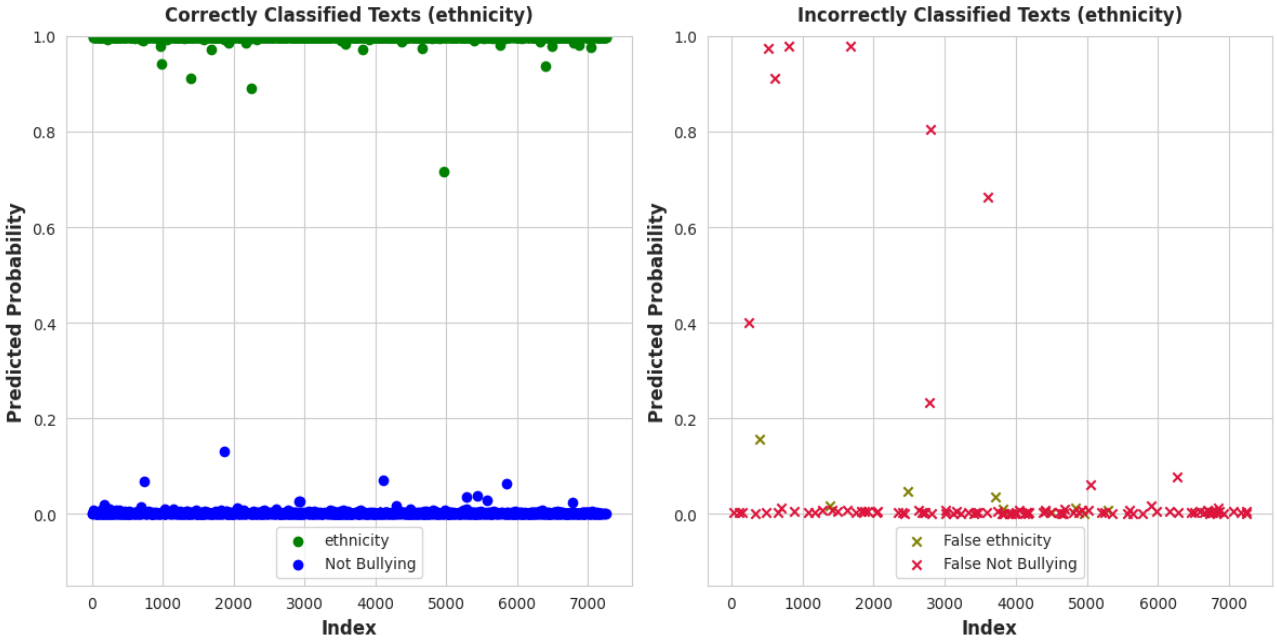


Рис. 4.14 – Розподіл коректно та некоректно класифікованих моделлю BERT текстових зразків щодо типу кіберзалякувань за етнічною приналежністю

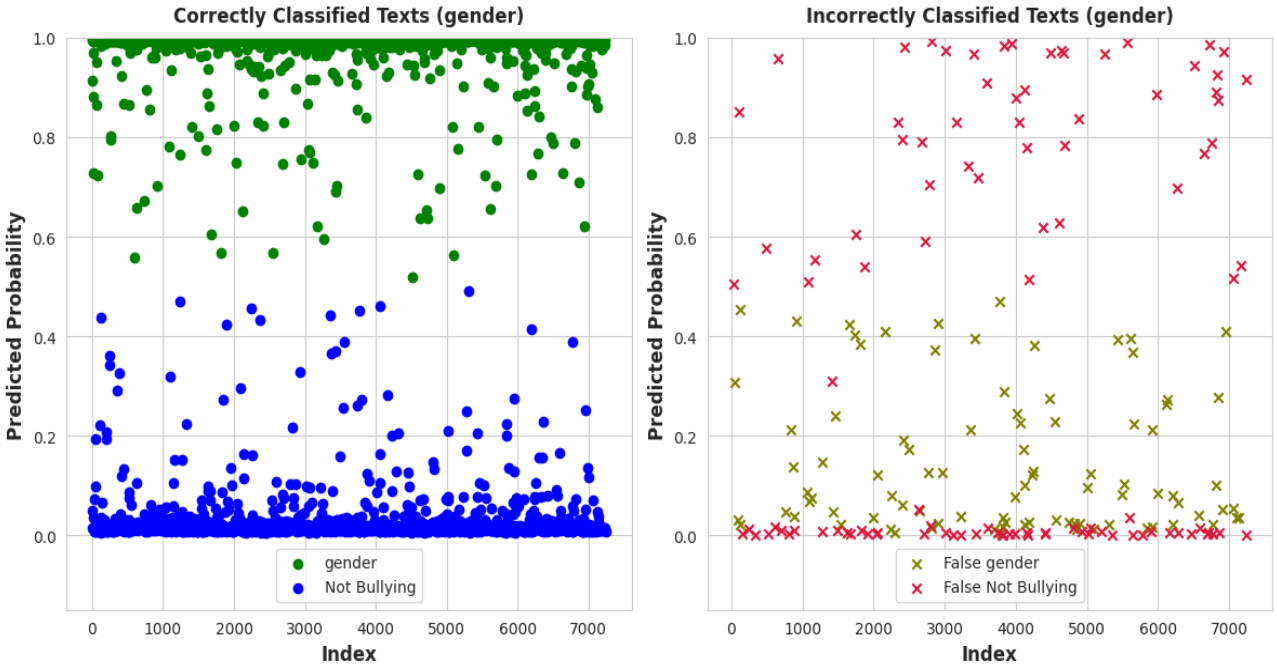


Рис. 4.15 – Розподіл коректно та некоректно класифікованих моделлю BERT текстових зразків щодо типу кіберзалякувань за гендером

Модель BERT правильно класифікувала за віковим кіберзалякуванням 1525 текстових зразків, неправильно – 27, за релігійним правильно класифіковані 1534, неправильно – 9, за етнічним правильно класифіковані – 1463, неправильно – 9, за гендерним правильно класифіковані – 1357, неправильно – 90. Як видно з наведених статистичних даних, найбільшу кількість хибних результатів модель надала при виявленні вікового кіберзалякування, а найкраще виявляла релігійний тип кіберзалякувань, проте це не впливає на загальні показники ефективності моделі BERT.

Для оцінювання моделі на різних комбінаціях даних, проведено кросвалідацію на вибірках випадкових даних з датасету (таблиця 4.8). У результаті проведеної крос-валідації отримані такі середні значення метрик: Accuracy – 94,08 %, Precision – 93,6 %, Recall – 93,54 %, а F1-міра – 93,56 %. Відхилення цих показників від макрометрик моделі становлять: Accuracy – 0,08 %, Precision – 0,6

%, Recall – 0,54 %, F1-міра – 0,56 %. Невеликі відхилення свідчать про високу відповідність результатів крос-валідації характеристикам моделі, особливо щодо точності. Це підтверджує стабільність та надійність моделі BERT у задачі виявлення та мультилейблової класифікації випадків кіберзалякувань у текстових повідомленнях.

Таблиця 4.8

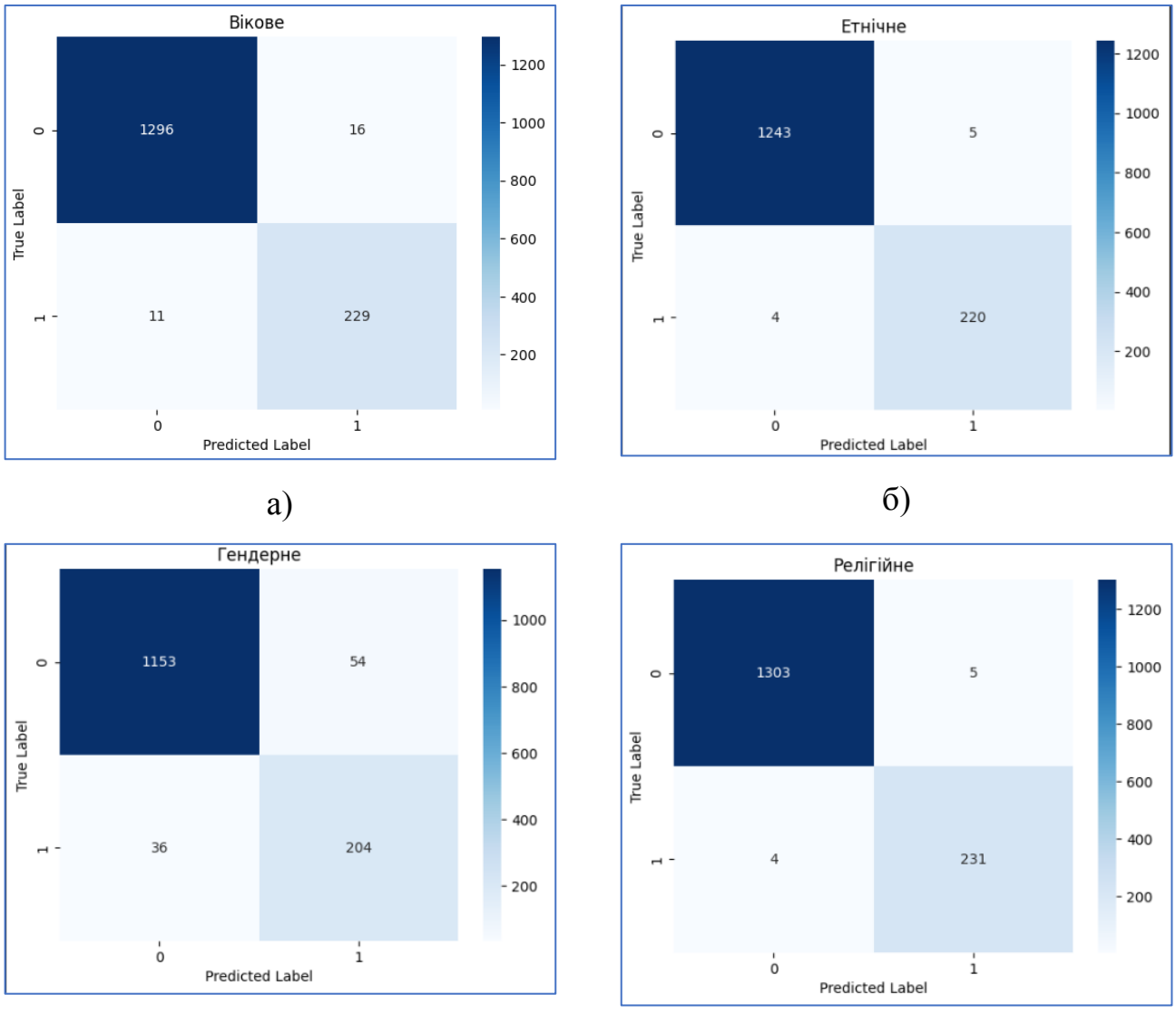
Показники метрик крос-валідації даних для моделі BERT

Вибірка	**Accuracy, %**	**Precision, %**	**Recall, %**	**F1, %**
Sample 1	93,8	93,4	93,2	93,3
Sample 2	94,0	93,5	93,6	93,5
Sample 3	94,2	93,6	93,4	93,5
Sample 4	94,1	93,8	93,7	93,7
Sample 5	94,3	93,7	93,8	93,8

Отже, для задачі бінарної класифікації кіберзалякувань було досліджено нейромережеві моделі BiLSTM та RNN з шаром LSTM. Для бінарної класифікації найкращі результати продемонструвала модель BiLSTM, досягнувши Accuracy 94 %, Precision – 94 %, Recall – 94 % та F_1-міри – 94 %. Для мультилейблової класифікації кіберзалякувань було досліджено моделі BiLSTM, LSTM, GRU, RoBERTa, BERT. Аналіз макро- та мікрометрик засвідчив, що найкращою серед них є модель BERT, яка досягла Accuracy 94 %, Precision – 93 %, Recall – 93 %, F_1 - міри – 93 %.

4.3. Експериментальне дослідження методу інтерпретації результатів виявлення кіберзалякувань

Для дослідження методу інтерпретації результатів виявлення кіберзалякувань було побудовано матрицю помилок мультилейблової класифікації кожного типу кіберзалякувань, яку наведено на рис. 4.16.



а) б)

в) г)

Рис. 4.16 – Матриці помилок для типів кіберзалякувань:
а) вікове; б) етнічне; в) гендерне; г) релігійне

Показники макрометрик навченої моделі BERT для мультилейблової класифікації типів кіберзалякувань отримали значення Accuracy – 94 %, Precision – 93 %, Recall – 93 %, F_1 Score – 93 %, що вказує на високу здатність моделі виявляти види кіберзалякувань у текстовому контенті.

Додатково було проведено експериментальне дослідження інтерпретаційної моделі на різних текстових зразках з метою дослідити, як описані у п. 2.2 обмеження методу впливають на результати інтерпретації.

Першим було досліджено обмеження, що стосується розміру вхідного тексту для аналізу на кіберзалякування. На вхід було подано текстовий зразок, що має менше, ніж 100 символів. На рис. 4.17 продемонстровано аналіз текстового зразка до 100 символів: «*Оце так новина, чого сюди приперся*? *Невже дома не сиділось*?».

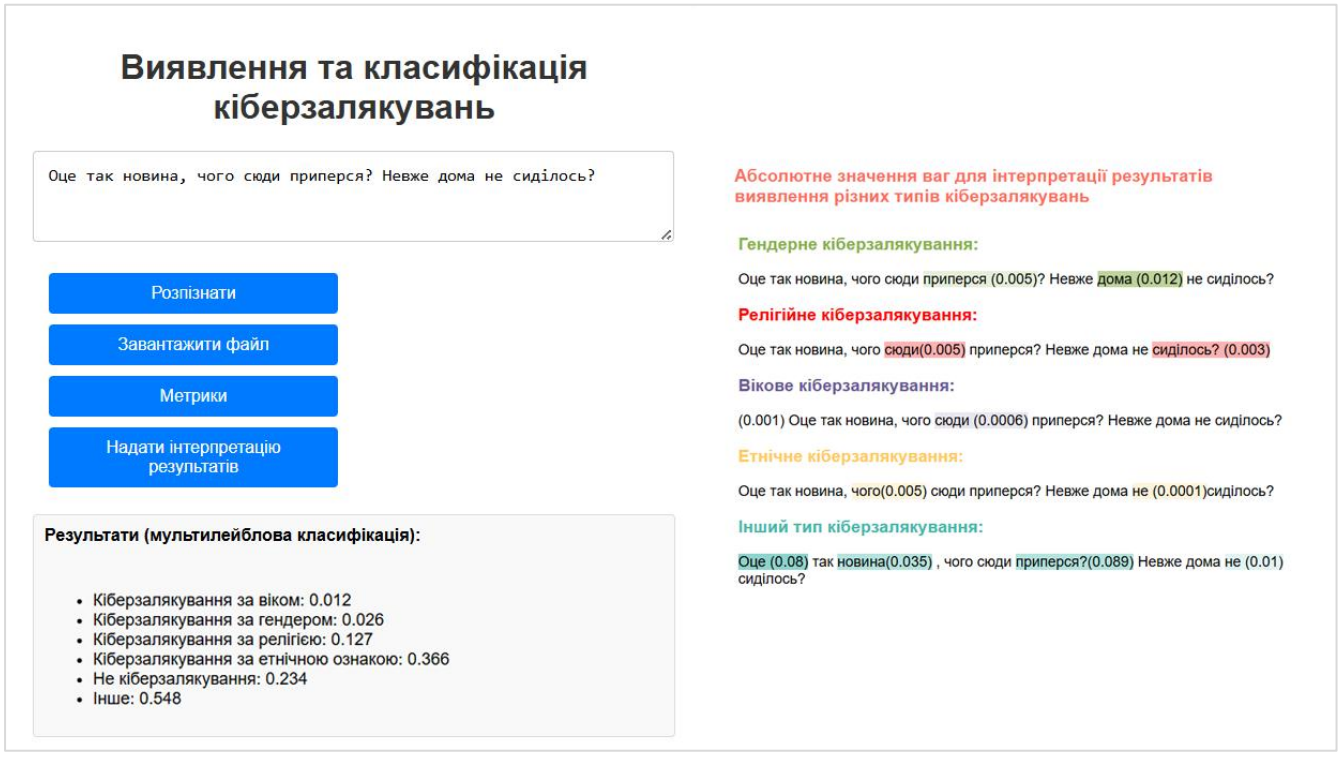


Рис. 4.17 – Результати аналізу текстового зразка до 100 символів

Як видно з рис. 4.17, на результат моделі BERT для мультилейблової класифікації більше вплинули загальні слова, які не є характерними для кожного з окремих типів кіберзалякувань. Це пов’язано з тим, що модель для мультилейблової класифікації навчалась на текстових даних у діапазоні 100–300 символів, а це означає, що розмір тексту до 100 символів значно впливає на точність класифікації кіберзалякувань у мультилейбловій моделі, оскільки зменшує обсяг контекстної інформації, необхідної для правильного визначення типу кіберзалякувань. У завданні виявлення та класифікації кіберзалякування граматичні конструкції та лексичні особливості відіграють ключову роль у розрізненні різних його типів. Скорочення тексту до 100 символів унеможливлює повноцінний аналіз складних мовних структур, що призводить до втрати важливих семантичних зв’язків між словами. Це є особливо критичним для мультилейблової класифікації, де правильне визначення типів кіберзалякувань залежить від взаємодії між лексичними одиницями.

Під час інтерпретації результатів LIME неправильно ідентифікує значущі слова, роблячи акцент на другорядних елементах, які насправді не визначають зміст кіберзалякування, як це видно на рис. 4.7. У результатах інтерпретації більшу вагу мають слова, які не є характерними для визначених типів кіберзалякувань. Більше того, кількість слів, які мають більшу вагу, є доволі обмеженою, що неповною мірою дозволяє пояснити отримані результати. Втрата контексту призводить до того, що навіть нейтральні або випадкові слова відзначені як ключові і мають великі ваги порівняно з іншими, а справді важливі характеристики залишаються поза увагою. Тому невелика довжина тексту призводить не лише до помилкової класифікації BERT, а й до спотворення пояснень, наданих LIME, що ускладнює інтерпретацію отриманих результатів та знижує достовірність проведеного аналізу.

Наступним проаналізовано текст, що є більшим за 300 символів: «*Ти абсолютно нікчемний, тебе ніхто не поважає, і всі тільки сміються з твоїх*

жалюгідних спроб щось зробити, бо в тебе ніколи нічого не вийде.. просто змирись з цим! Кожен твій пост – це повний сором, навіть діти могли б краще написати. Може, варто просто зникнути з інтернету? Ніхто не хоче бачити такі дурниці. Ти навіть не усвідомлюєш, наскільки смішно виглядаєш!» (рис. 4.18).

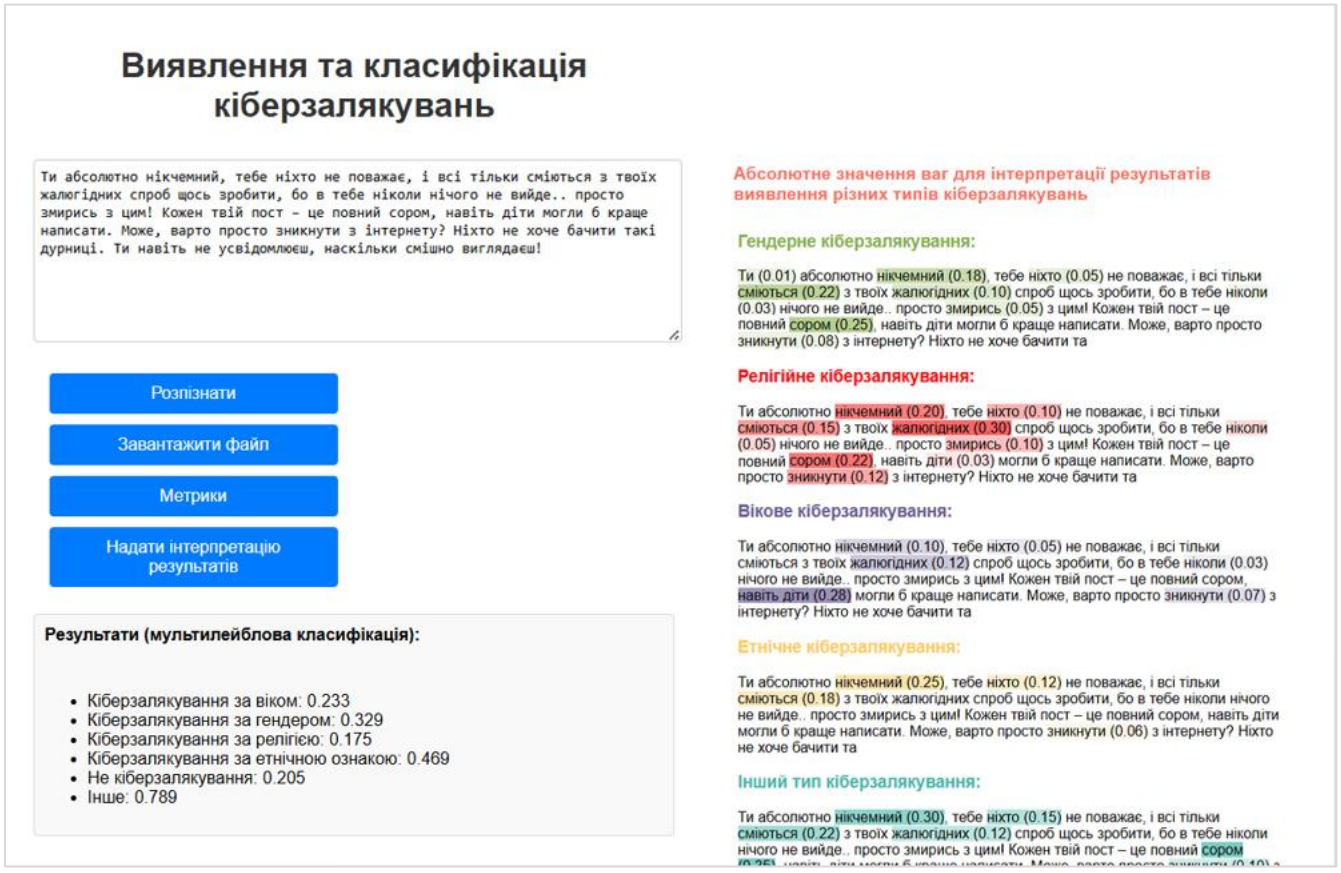


Рис. 4.18 – Результати аналізу текстового зразка понад 300 символів

На рис. 4.18 видно, що текст, який інтерпретує модель LIME для кожного типу кіберзалякування, відображений неповністю. Оскільки модель BERT для мультилейблової класифікації була навчена на текстових зразках, що не перевищують 300 символів, то будь-який текстовий контент, довший за встановлений ліміт, підлягає скороченню до цього розміру. В свою чергу, це спричиняє втрату частини тексту, який міститиме контекст для більш коректного

визначення типу кіберзалякувань. Метод інтерпретації LIME також зазнає негативного впливу через вкорочення вхідного повідомлення. Наприклад, якщо втрачається частина речення, яка містить контекст кіберзалякування за етнічною чи релігійною ознакою, LIME може помилково вважати, що кіберзалякування відбувається за іншим критерієм або взагалі відсутнє. Тому це призводить до ситуацій, коли модель BERT неправильно розпізнає тип кіберзалякування, а пояснення LIME акцентує увагу на непринципових частинах текстового зразка, роблячи результати менш достовірними, як показано на рис. 4.18.

Також був проаналізований на кіберзалякування та його типи текстовий зразок, що містить суржик та сленг «*Діду, шо ти там опять пишеш? Вже б давно на пенсії відпочивав, а не позорився тут. Та кому цікаві твої допотопні думки? Всі нормальні люди вже давно пішли вперед, а ти як той динозавр, шо вимерти забув!*» (рис. 4.19).

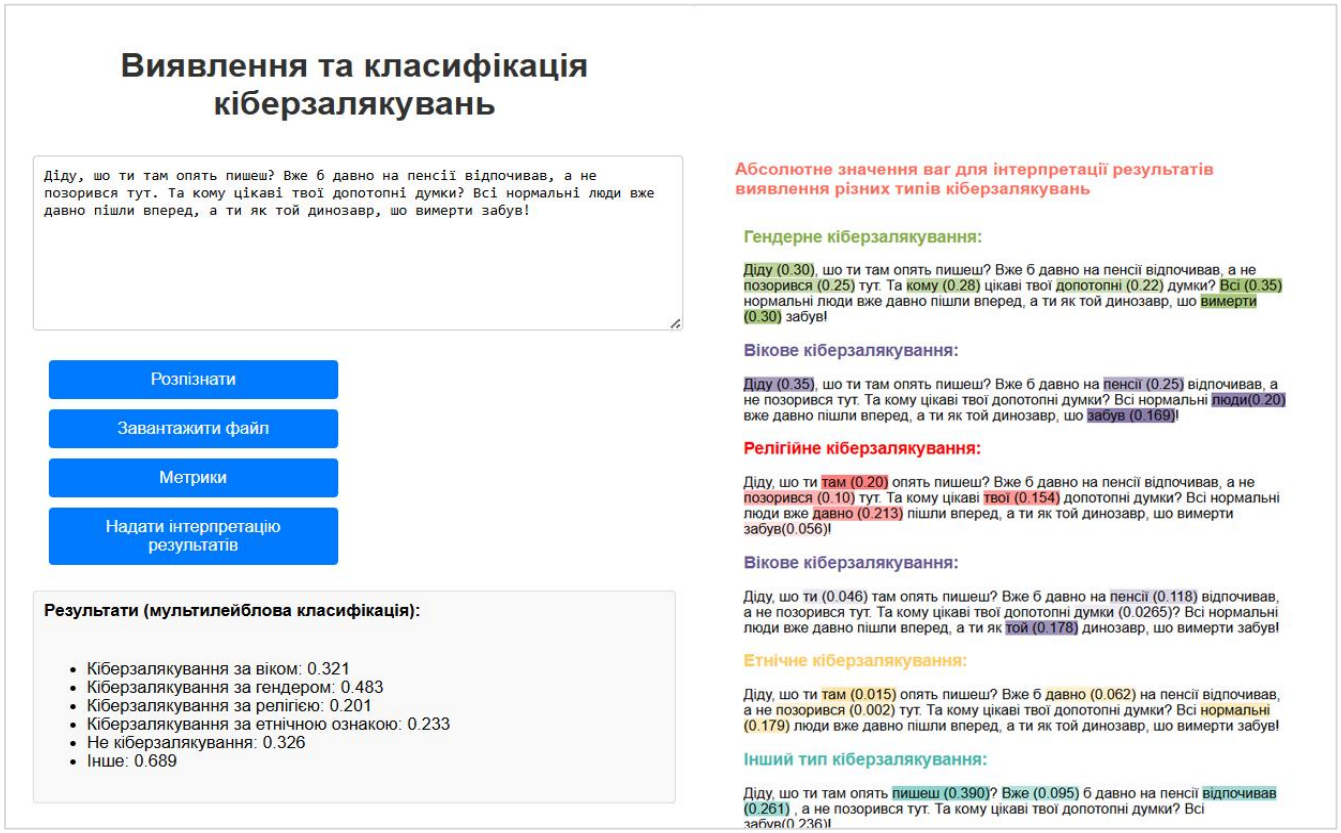


Рис. 4.19 – Результати аналізу текстового зразка, що містить суржик та сленг

Автоматизований переклад, який використовувався для перекладу датасетів для навчання нейромережевих моделей, не завжди здатний правильно розпізнати контекст і стилістичні особливості українських виразів.

Проблема класифікації текстів, що містять суржик та сленг, при використанні перекладених датасетів для навчання моделей виникає через особливості автоматичного машинного перекладу, який не завжди коректно відтворює мовні конструкції. В англійській мові відсутнє явище суржику, оскільки вона є мовно уніфікованою і не містить значного змішування з іншими мовами на рівні побутового чи офіційного використання. Українська мова, навпаки, в силу історичних та соціолінгвістичних факторів, має велику кількість запозичень, які часто не усвідомлюються носіями, в тому числі при спілкуванні онлайн.

При перекладі англомовних текстів, що містять нейтральні або розмовні вирази, система перекладу може інтерпретувати їх через найбільш поширені відповідники, що особливо актуально для сленгу, лайливих висловів, що характерні для кіберзалякувань. Автоматичний переклад також часто спрощує та змінює синтаксичні конструкції, використовуючи слова, які частіше зустрічаються в українському мовленні. Тому це негативно впливає на якість класифікації, оскільки нейромережа, навчена на чистій українській мові або стандартних перекладених текстах, неправильно інтерпретує суржикові та сленгові конструкції.

Як показує проведений аналіз текстового зразка на рис. 4.19, використання суржику та сленгу спричинило неправильне розпізнавання типів кіберзалякувань. У наведеному тексті міститься значна кількість лексичних конструкцій, які є характерними для вікового кіберзалякування. Проте їх використання у формі

суржику та сленгу змінює семантичне навантаження та ускладнює коректне розпізнавання нейромережевою моделлю. Оскільки модель була навчена на текстах, що перекладені українською мовою, вона сприймає суржик та сленг як менш релевантні до вікового кіберзалякування в цьому випадку, а це спричиняє зниження точності цього класу. Водночас, через неоднозначність аналізу, модель відносить такі фрагменти тексту до узагальненого класу «інші типи кіберзалякувань», що й призводить до підвищеної точності саме для цього типу.

Окрім цього LIME також неправильно трактує слова, що є суржиком або сленгом. У випадку текстів із суржиком та сленгом модель сприймає такі слова як незвичні або такі, що не мають достатнього представлення в навчальних даних. Тому LIME надає більшу вагу для інших окремих слів, які не є ключовими для конкретних окремих типів кіберзалякувань. В результаті інтерпретація є хибною, оскільки вагомі слова для вікового кіберзалякування мають меншу вагу, а інші слова у тексті отримують більшу вагу, що знову ж таки сприяє збільшенню впевненості у класі «інші типи кіберзалякувань».

Отже, з проведених досліджень можна зробити висновки, що обмеження, пов’язані з розміром тексту, наявністю суржику та сленгу впливають на точність класифікації типів кіберзалякувань у текстовому контенті та на інтерпретацію результатів методом LIME. Тексти менші за 100 символів не забезпечують достатнього контексту для коректної класифікації типів кіберзалякувань, а тексти понад 300 символів призводять до втрати частини контенту через обмеження моделі. Наявність суржику та сленгу у текстах додає ще одне суттєве обмеження для моделі класифікації та інтерпретації отриманих результатів: суржик та сленг змінюють семантику та призводять до неправильної інтерпретації контексту. Тому ці обмеження ускладнюють процес виявлення та класифікації кіберзалякувань.

4.4. Висновки до розділу 4

Для дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості була створена підсистема, яка використовує моделі машинного та глибокого навчання для класифікації текстів за різними етичними аспектами – віком, статтю, релігією. У дослідженні SVM використовувався для класифікації зразків текстів у вибірці за віковим етичним аспектом, LSTM – для гендеру, BERT – для релігійного аспекту.

У результаті практичного застосування розробленого методу встановлено, що датасет для виявлення кіберзалякувань не є репрезентативним порівняно з даними демографічної статистики населення України, тому з використанням розробленого методу було розв’язано задачу багатокритеріальної оптимізації та приведено датасет до репрезентативного вигляду з врахуванням вікового і гендерного етичного аспектів. Отримані відхилення розподілів за етичними аспектами скоригованого датасету від цільових пропорцій склали: мінімальне – 0,00 %, максимальне – 0,04 %, середнє – 0,02 %.

Проведене дослідження доводить, що розроблений метод дозволяє виконувати оцінку репрезентативності датасетів та коригування їх репрезентативності відповідно до різних етичних аспектів принципу справедливості FATE. Встановлено, що підвищення рівню репрезентативності датасетів за етичними аспектами впливає на якість навчання класифікаторів для виявлення кіберзалякувань та їх типів.

Дослідження методу виявлення кіберзалякувань у текстовому контенті спрямоване на аналіз його можливостей аналізувати текстові повідомлення та визначати загальний рівень прояву кіберзалякувань, а також виконувати мультилейблову класифікацію, надаючи окремі показники для оцінювання проявів

різних типів кіберзалякувань, таких як вікові, релігійні, етнічні, гендерні залякування тощо.

У завданні бінарної класифікації кращі результати показала модель BiLSTM з показниками Accuracy – 94 %, Precision – 94 %, Recall – 94 %, F_1-міри – 94 %. Для завдання мультилейблової класифікації модель BERT показала значення макрометрик Accuracy – 94 %, Precision – 93 %, Recall – 93 %, F_1-міри – 93 %. Дослідивши значення мікрометрик для кожного з типів кіберзалякувань, встановлено, що модель BERT показує більш стабільні показники для класифікації різних типів кіберзалякувань, аніж інші досліджені моделі (BiLSTM, LSTM, GRU, RoBERTa).

У результаті дослідження методу інтерпретації результатів виявлення кіберзалякувань експериментально встановлено, що обмеження розміру тексту та наявність суржику і сленгу впливають на точність класифікації типів кіберзалякувань. Встановлено, що розмір тексту менше 100 символів обмежує контекст, що знижує точність класифікації. Суржик і сленг змінюють значення висловлювань, ускладнюючи правильне розпізнавання контексту, що призводить до помилок у класифікації, оскільки модель не була навчена на таких мовних конструкціях. У підсумку, інтерпретація результатів виявлення типів кіберзалякувань виявляється некоректною, оскільки ключові слова, характерні для одного типу кіберзалякування, отримують нижчу вагу, тоді як інші слова в тексті мають вищу значущість.

ВИСНОВКИ

У результаті виконання дисертаційного дослідження було розв’язано актуальну науково-прикладну задачу виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту. Відповідно до мети дослідження, яка полягала у підвищенні точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень, отримано такі науково-практичні результати:

1. На основі проведеного аналізу методів, засобів і технологій для автоматизованого виявлення кіберзалякувань у текстовому контенті визначено потребу у розробці нових методів, що підвищуватимуть якість виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень. Аналіз підтвердив актуальність застосування засобів штучного інтелекту для виявлення кіберзалякувань у текстовому контенті, проте виявлено ряд суперечностей. Для виявлення кіберзалякувань використовуються моделі, що навчені на датасетах, які не є репрезентативними, що призводить до дискримінації соціальних груп моделями та знижує справедливість отриманих результатів. Також відомі підходи не надають інтерпретації результатів виявлення типів кіберзалякувань для мультилейблової класифікації.

2. 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. Метод включає три кроки: перевірку коректності елементів датасету, аналіз репрезентативності за етичними аспектами, репрезентативне коригування датасету. Для аналізу репрезентативності за етичними аспектами використано моделі машинного та глибокого навчання, що показали кращі

показники точності під час експериментального дослідження. Підвищення якості навчання класифікаторів для виявлення кіберзалякувань полягає у формуванні репрезентативного та недискримінаційного за FATE-принципом справедливості датасету.

3. Розроблено новий метод виявлення кіберзалякувань у текстовому контенті, який відрізняється від існуючих двоетапним виявленням кіберзалякувань, що полягає у нейромережевій ідентифікації наявності кіберзалякувань і подальшій нейромережевій мультилейбловій класифікації окремих типів кіберзалякувань, що дало можливість підвищити точність та якість виявлення кіберзалякувань. Метод включає три кроки: попередню обробку тексту для аналізу, аналіз тексту на наявність кіберзалякувань, мультилейблову класифікацію типів кіберзалякувань у випадку, якщо кіберзалякування було виявлене на попередньому кроці. Підвищення якості виявлення кіберзалякувань полягає у застосуванні двоетапної перевірки – спочатку виявлення кіберзалякування, а потім визначення наявних типів кіберзалякувань у текстовому контенті. За результатами досліджень, для випадку бінарної класифікації кращі результати показала нейромережева модель BiLSTM з показниками Accuracy – 94 %, Precision – 94 %, Recall – 94 %, F_1-міри – 94 %; для випадку мультилейблової класифікації кращі результати показала нейромережева модель BERT з показниками макрометрик Accuracy – 94 %, Precision – 93 %, Recall – 93 %, F_1-міри – 93 %. Виявлення кіберзалякувань запропонованим у дослідженні методом має показник Accuracy щонайменше на 0,8 % вищий, ніж у відомих підходах.

4. Удосконалено метод інтерпретації результатів виявлення кіберзалякувань, який відрізняється від існуючих можливістю надавати візуальні пояснення для мультилейблової класифікації виявлених типів кіберзалякувань в альтернативних поданнях. Метод включає три етапи: попередню обробку тексту для аналізу, мультилейблову класифікацію типів кіберзалякувань, візуалізацію впливу ознак для кожного типу. Підвищення якості виявлення кіберзалякувань

таким чином здійснюється шляхом використання мультилейблового класифікатора для класифікації типів кіберзалякувань на основі нейромережевої архітектури трансформер та інтерпретаційної моделі. Метод забезпечує візуальну інтерпретацію результатів у вигляді колірного представлення ваг слів, що вплинули найбільше на рішення моделі, а також у вигляді діаграм впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування та середнього значення важливості топ 10 слів для всіх класів.

5. Запропоновані у дисертаційному дослідженні методи реалізовано у вигляді інтелектуальної інформаційної системи для виявлення та класифікації кіберзалякувань у текстовому контенті, що на основі навчених нейромережевих моделей BiLSTM та BERT, а також розробленого методу оцінювання і коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечує недискримінацію за віковою, етнічною, гендерною, релігійною приналежністю та методу інтерпретації результатів виявлення кіберзалякувань, дозволяють підвищити якість їхніх виявлення у текстовому контенті. Проведені експериментальні дослідження та порівняння дозволили підтвердити підвищення точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень, що і було метою дисертаційного дослідження.

Практичне значення отриманих результатів полягає у доведенні теоретичних результатів дисертаційної роботи та розробці інтелектуальної інформаційної системи, що використовує розроблені методи оцінювання та коригування репрезентативності датасету, виявлення та класифікації кіберзалякувань, а також інтерпретації результатів їх виявлення, і дозволяє підвищити точність та якість виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту й візуально пояснювати прийняті рішення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Собко О. В. Нейромережевий пошук і класифікація кіберзалякувань у текстових повідомленнях. *Information Technology: Computer Science, Software Engineering and Cybersecurity.* 2024. № 4. С. 197–205. URL: https://doi.org/10.32782/IT/2024-4-23 (дата звернення: 19.03.2025).

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 (дата звернення: 19.03.2025).

3. Собко О. В. Метод інтелектуального виявлення кіберзалякувань у текстовому контенті. *Розвитки інформаційно-керуючих систем та технологій : монографія* / Н. Аксак, Д. Антонов та ін. ; під наук. ред. проф. В. Вичужаніна. Львів–Торунь : Lina-Pres, 2024. С. 267–287. URL: http://catalog.liha-pres.eu/index.php/liha-pres/catalog/view/319/9254/20840-1 (дата звернення: 19.03.2025).

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 (дата звернення: 19.03.2025).

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 (дата звернення: 19.03.2025).

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 (дата звернення: 19.03.2025).

7. Method for Neural Network Cyberbullying Detection in Text Content With Visual Analytic / I. Krak, O. Sobko, M. Molchanova, I. Tymofiiev, O. Mazurets, O. Barmak. *CEUR Workshop Proceedings*, 2025, vol. 3917, pp. 298–309. URL: https://ceur-ws.org/Vol-3917/paper57.pdf (дата звернення: 19.03.2025).

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 (дата звернення: 19.03.2025).

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 (дата звернення: 19.03.2025).

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. № 133211 Україна. Твір наукового характеру «Методи визначення та класифікації кіберзалякувань у текстовому контенті з використанням засобів штучного інтелекту» / О. В. Собко. 2025.

12. Закон України «Про основні засади забезпечення кібербезпеки України» № 2145-VIII від 05.10.2017, *Верховна Рада України*. URL: https://zakon.rada.gov.ua/laws/show/2145-19#Text (дата звернення: 23.03.2025).

13. Кодекс України про адміністративні правопорушення, *Верховна Рада України*. URL: https://zakon.rada.gov.ua/laws/show/80731-10#n4218 (дата звернення: 23.03.2025).

14. Наказ «Про створення безпечного освітнього середовища в закладі освіти та попередження і протидію булінгу (цькуванню)», *Міністерство освіти і науки України.* URL: https://mon.gov.ua/npa/pro-stvorennya-bezpechnogo-osvitnogo-

seredovisha-v-zakladi-osviti-ta-poperedzhennya-i-protidiyi-bulingu-ckuvannyu (дата звернення: 23.03.2025).

15. Cyberbullying intervention and prevention programmes in Primary Education (6 to 12 years): A systematic review / M. Chicote-Beato et al. *Aggression and Violent Behavior*. 2024. P. 101938. URL: https://doi.org/10.1016/j.avb.2024.101938 (дата звернення: 24.03.2025).

16. Teng T. H., Varathan K. D., Crestani F. A Comprehensive Review of Cyberbullying-related Content Classification In Online Social Media. *Expert Systems with Applications*. 2023. P. 122644. URL: https://doi.org/10.1016/j.eswa.2023.122644 (дата звернення: 24.03.2025).

17. Global Research Trends on Cyberbullying: A Bibliometric Study / A. Singh et al. *Computers in Human Behavior Reports*. 2024. P. 100499. URL: https://doi.org/10.1016/j.chbr.2024.100499 (дата звернення: 24.03.2025).

18. 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 (дата звернення: 24.03.2025).

19. Psychological Impacts, Prevention Strategies, and Intervention Approaches Across Age Groups / J. S. Kushwah et al. *Change Dynamics in Healthcare, Technological Innovations, and Complex Scenarios*. 2024. P. 89–109. URL: https://doi.org/10.4018/979-8-3693-3555-0.ch005 (дата звернення: 24.03.2025).

20. One in six school-aged children experiences cyberbullying, finds new WHO/Europe study. *World Health Organization (WHO).* URL: https://www.who.int/europe/news/item/27-03-2024-one-in-six-school-aged-children-experiences-cyberbullying--finds-new-who-europe-study (дата звернення: 23.03.2025).

21. Mental Health and Bias-Based Bullying and Cyberbullying Victimization Among Young Adults with Intersectional Identities / R. Y. Feng et al. *International Journal of Bullying Prevention*. 2024. URL: https://doi.org/10.1007/s42380-024-00260-7 (дата звернення: 23.03.2025).

22. Aggarwal A., A Study of the Factors that Influence Cyber Bullying-*Perspectives from Bullies*, Int. J. Eng. Manag. Res. 14.3 (2024) 1–5. URL: https://doi.org/10.5281/zenodo.11448952 (дата звернення: 23.03.2025).

23. UNICEF. Cyberbullying. URL: https://www.unicef.org/ukraine/cyberbulling (дата звернення: 23.03.2025).

24. A Cross-National Perspective of Prejudice-Based Cyberbullying and Cybervictimisation / B. M. Dinić et al. *International Perspectives on Migration, Bullying, and School*. London, 2024. P. 148–165. URL: https://doi.org/10.4324/9781003439202-9 (дата звернення: 23.03.2025).

25. Civila S. Cyberbullying. *Comprehensive Sexuality Education for Gender-Based Violence Prevention*. 2024. P. 229–245. URL: https://doi.org/10.4018/979-8-3693-2053-2.ch013 (дата звернення: 23.03.2025).

26. Casas F. Age Discrimination. *Encyclopedia of Quality of Life and Well-Being Research*. Cham, 2023. P. 118–121. URL: https://doi.org/10.1007/978-3-031-17299-1_48 (дата звернення: 23.03.2025).

27. Lee H. Lived Religion in Religious Vaccine Exemptions. *Perspectives in Biology and Medicine*. 2024. Vol. 67, no. 1. P. 96–113. URL: https://doi.org/10.1353/pbm.2024.a919713 (дата звернення: 23.03.2025).

28. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity / C. Emmery et al. *Language Resources and Evaluation*. 2020. URL: https://doi.org/10.1007/s10579-020-09509-1 (дата звернення: 23.03.2025).

29. Home. *Cyber.bullyingstop*. URL: https://cyber.bullyingstop.org.ua (дата звернення: 23.03.2025).

30. Протидія булінгу. *MOH.* URL: https://mon.gov.ua/tag/protidiya-bulingu?&type=all&tag=protidiya-bulingu (дата звернення: 23.03.2025).

31. Антибулінг. *AIKOM.* URL: https://aikom.iea.gov.ua/bullying/help (дата звернення: 23.03.2025).

32. Yengejeh A. A., Combating Cyberbullying on Social Media: A Machine Learning Approach with Text Analysis on Twitter. *Data Science and Data Mining*, 15. 2024. URL: https://core.ac.uk/download/pdf/599808315.pdf (дата звернення: 23.03.2025).

33. Teng T. H., Varathan K. D. Cyberbullying Detection in Social Networks: A Comparison between Machine Learning and Transfer Learning Approaches. *IEEE Access*. 2023. P. 1. URL: https://doi.org/10.1109/access.2023.3275130 (дата звернення: 23.03.2025).

34. 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. URL: https://doi.org/10.48084/etasr.7621 (дата звернення: 23.03.2025).

35. AlDahoul N., Tan M. J. T., Kasireddy H. R., Zaki Y., Advancing Content Moderation: Evaluating Large Language Models for Detecting Sensitive Content Across Text, Images, and Videos, *arXiv preprint arXiv:2411.17123* (2024). URL: https://arxiv.org/abs/2411.17123 (дата звернення: 23.03.2025).

36. Comparison of Machine Learning and Deep Learning Models for Detecting Cyberbullying / K. A. Lo et al. *2024 International Visualization, Informatics and Technology Conference (IVIT)*, Kuala Lumpur, Malaysia, 7–8 August 2024. 2024. P. 138–144. URL: https://doi.org/10.1109/ivit62102.2024.10692892 (дата звернення: 23.03.2025).

37. Liang X. The Cause and Influence of Cyberbullying. *Journal of Education, Humanities and Social Sciences*. 2024. Vol. 26. P. 661–668. URL: https://doi.org/10.54097/zp49e018 (дата звернення: 23.03.2025).

38. Cyberbullying: Research Challenges and Opportunities / S. Nizam et al. *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 9–10 February 2024. 2024. URL: https://doi.org/10.1109/ic2pct60090.2024.10486805 (дата звернення: 23.03.2025).

39. E. F. C., Strategies for Dealing with Bullying and Cyberbullying, *Practici Integrative pentru Prevenirea și Combaterea Fenomenului de Bullying în Organizațiile Școlare* 83 (2024). URL: https://ibn.idsi.md/sites/default/files/imag_file/Practici_integrative_pu_prevenirea_si_combaterea_fenomen._bullying_16.02.2024.pdf#page=84 (дата звернення: 23.03.2025).

40. Shah M., Sureja N. A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions. *Archives of Computational Methods in Engineering*. 2024. URL: https://doi.org/10.1007/s11831-024-10134-2 (дата звернення: 23.03.2025).

41. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods / T. P. Pagano et al. *Big Data and Cognitive Computing*. 2023. Vol. 7, no. 1. P. 15. URL: https://doi.org/10.3390/bdcc7010015 (дата звернення: 23.03.2025).

42. Cyberbullying Classification. *Kaggle.com.* 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download (дата звернення: 23.03.2025).

43. CyberBullying Detection Dataset. *Kaggle.com.* 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (дата звернення: 23.03.2025).

44. Cyberbullying Detection. *Kaggle.com.* 2023. URL: https://www.kaggle.com/datasets/gbiamgaurav/cyberbullying-detection (дата звернення: 23.03.2025).

45. Cyberbullying Dataset. *Kaggle.com.* 2023. URL: https://www.kaggle.com/datasets/ashiqnazir/cbtweets (дата звернення: 23.03.2025).

46. Elazar Y., Goldberg Y. Adversarial Removal of Demographic Attributes from Text Data. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Stroudsburg, PA, USA, 2018. URL: https://doi.org/10.18653/v1/d18-1002 (дата звернення: 23.03.2025).

47. Uncurated Image-Text Datasets: Shedding Light on Demographic Bias / N. Garcia et al. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 17–24 June 2023. 2023. URL: https://doi.org/10.1109/cvpr52729.2023.00672 (дата звернення: 23.03.2025).

48. Impact of Annotator Demographics on Sentiment Dataset Labeling / Y. Ding et al. *Proceedings of the ACM on Human-Computer Interaction*. 2022. Vol. 6, CSCW2. P. 1–22. URL: https://doi.org/10.1145/3555632 (дата звернення: 23.03.2025).

49. Toward Responsible Artificial Intelligence in Long-Term Care: A Scoping Review on Practical Approaches / D. R. M. Lukkien et al. *The Gerontologist*. 2021. URL: https://doi.org/10.1093/geront/gnab180 (дата звернення: 23.03.2025).

50. Memarian B., Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence*. 2023. P. 100152. URL: https://doi.org/10.1016/j.caeai.2023.100152 (дата звернення: 23.03.2025).

51. Geleta R. R., Exploring the Role of AI and XAI in Hate Speech Detection on Social Media: A Study on User Trust, 2023. URL: https://epub.jku.at/obvulihs/content/titleinfo/8885187 (дата звернення: 23.03.2025).

52. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions / L. Longo et al. *Information Fusion*. 2024. P. 102301. URL: https://doi.org/10.1016/j.inffus.2024.102301 (дата звернення: 24.03.2025).

53. He G., Aishwarya N., Gadiraju U. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. *IUI '25: 30th International Conference on Intelligent User Interfaces*, Cagliari Italy. New York, NY, USA, 2025. P. 907–924. URL: https://doi.org/10.1145/3708359.3712133 (дата звернення: 24.03.2025).

54. Thalpage N. Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems. *Journal of Digital Art & Humanities*.

2023. Vol. 4, no. 1. P. 31–36. URL: https://doi.org/10.33847/2712-8148.4.1_4 (дата звернення: 23.03.2025).

55. Akhai S. From Black Boxes to Transparent Machines: The Quest for Explainable AI. *SSRN Electronic Journal*. 2023. URL: https://doi.org/10.2139/ssrn.4390887 (дата звернення: 23.03.2025).

56. Saeed W., Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*. 2023. Vol. 263. P. 110273. URL: https://doi.org/10.1016/j.knosys.2023.110273 (дата звернення: 23.03.2025).

57. Liu Z., He K., A Decade's Battle on Dataset Bias: Are We There Yet?, *arXiv preprint arXiv:2403.08632* (2024). URL: https://doi.org/10.48550/arXiv.2403.08632 (дата звернення: 23.03.2025).

58. Potential Applications of Explainable Artificial Intelligence to Actuarial Problems / C. Lozano-Murcia et al. *Mathematics*. 2024. Vol. 12, no. 5. P. 635. URL: https://doi.org/10.3390/math12050635 (дата звернення: 23.03.2025).

59. Mahto M. K., Explainable Artificial Intelligence: Fundamentals, Approaches, Challenges, XAI Evaluation, and Validation, *in: Explainable Artificial Intelligence for Autonomous Vehicles*, CRC Press, 2025, pp. 25–49. URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003502432-2/explainable-artificial-intelligence-manoj-kumar-mahto (дата звернення: 23.03.2025).

60. Xu Q. et al., Applications of Explainable AI in Natural Language Processing, *Glob. Acad. Front*. 2.3 (2024) 51–64. URL: https://doi.org/10.5281/zenodo.12684705 (дата звернення: 23.03.2025).

61. Rane N. L., Paramesha M. Explainable Artificial Intelligence (XAI) as a foundation for trustworthy artificial intelligence. *Trustworthy Artificial Intelligence in Industry and Society*. 2024. URL: https://doi.org/10.70593/978-81-981367-4-9_1 (дата звернення: 23.03.2025).

62. A methodology to compare XAI explanations on natural language processing / G. Jouis et al. *Explainable Deep Learning AI*. 2023. P. 191–216.

URL: https://doi.org/10.1016/b978-0-32-396098-4.00016-8 (дата звернення: 23.03.2025).

63. Detecting and Understanding Cyberbullying in Bengali: An Explainable AI Approach / M. K. Syfullah et al. *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, Cox's Bazar, Bangladesh, 25–26 September 2024. 2024. P. 1–7. URL: https://doi.org/10.1109/compas60761.2024.10797170 (дата звернення: 24.03.2025).

64. Datasets: A Community Library for Natural Language Processing / Q. Lhoest et al. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online and Punta Cana, Dominican Republic. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.emnlp-demo.21 (дата звернення: 23.03.2025).

65. Mohammed S., Ehrlinger L., Harmouch H., Naumann F., Srivastava D., Data Quality Assessment: Challenges and Opportunities, *arXiv preprint arXiv:2403.00526* (2024). URL: https://arxiv.org/abs/2403.00526 (дата звернення: 23.03.2025).

66. NLPositionality: Characterizing Design Biases of Datasets and Models / S. Santy et al. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.acl-long.505 (дата звернення: 23.03.2025).

67. Structural Alignment Method of Conceptual Categories of Ontology and Formalized Domain / E. Manziuk, I. Krak, O. Barmak, O. Mazurets, V. Kuznetsov, O. Pylypiak. CEUR Workshop Proceedings, 3003, 2021, pp. 11–22. URL: https://ceur-ws.org/Vol-3003/paper2.pdf (дата звернення: 23.03.2025).

68. Data Representativeness in Accessibility Datasets: A Meta-Analysis / R. Kamikubo et al. *ASSETS '22: The 24th International ACM SIGACCESS Conference on Computers and Accessibility*, Athens Greece. New York, NY, USA, 2022. URL: https://doi.org/10.1145/3517428.3544826 (дата звернення: 23.03.2025).

69. Clemmensen L. H., Kjærsgaard R. D., Data Representativity for Machine Learning and AI Systems, *arXiv preprint arXiv:2203.04706* (2022). URL: https://doi.org/10.48550/arXiv.2203.04706 (дата звернення: 23.03.2025).

70. Dablain D., Krawczyk B., Chawla N. Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. *Discover Data*. 2024. Vol. 2, no. 1. URL: https://doi.org/10.1007/s44248-024-00007-1 (дата звернення: 23.03.2025).

71. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias / R. K. E. Bellamy et al. *IBM Journal of Research and Development*. 2019. Vol. 63, no. 4/5. P. 4:1–4:15. URL: https://doi.org/10.1147/jrd.2019.2942287 (дата звернення: 23.03.2025).

72. Benchmarking Intersectional Biases in NLP / J. Lalor et al. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Stroudsburg, PA, USA, 2022. URL: https://doi.org/10.18653/v1/2022.naacl-main.263 (дата звернення: 23.03.2025).

73. Raza S., Reji D. J., Ding C. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*. 2022. URL: https://doi.org/10.1007/s41060-022-00359-4 (дата звернення: 23.03.2025).

74. Chen H., Ji Y., Evans D. Addressing Both Statistical and Causal Gender Fairness in NLP Models. *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico. Stroudsburg, PA, USA, 2024. URL: https://doi.org/10.18653/v1/2024.findings-naacl.38 (дата звернення: 23.03.2025).

75. Zhou K. et al., Fairpriori: Improving Biased Subgroup Discovery for Deep Neural Network Fairness, *arXiv preprint arXiv:2407.01595* (2024). URL: *https://doi.org/10.48550/arXiv.2407.01595* (дата звернення: 23.03.2025).

76. Evans A. S., Moniz H., Coheur L., A Study on Bias Detection and Classification in Natural Language Processing, *arXiv preprint arXiv:2407.01595* (2024). URL: https://doi.org/10.48550/arXiv.2407.01595 (дата звернення: 23.03.2025).

77. Attribute-specific Cyberbullying Detection Using Artificial Intelligence / A. Orelaja et al. *Journal of Electronic & Information Systems*. 2024. Vol. 6, no. 1. P. 10–21. URL: https://doi.org/10.30564/jeis.v6i1.6206 (дата звернення: 23.03.2025).

78. Кібербулінг – що це та як це зупинити? *Unicef*. URL: https://www.unicef.org/ukraine/cyberbulling (дата звернення: 23.03.2025).

79. Using Sarcasm to Improve Cyberbullying Detection / X. Guo, S. Gauch. In: *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, Torino, Italia. ELRA and ICCL, 2024, pp. 52–59. URL: https://aclanthology.org/2024.trac-1.7/ (дата звернення: 23.03.2025).

80. Ejaz N., Razi F., Choudhury S. Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior*. 2023. P. 108123. URL: https://doi.org/10.1016/j.chb.2023.108123 (дата звернення: 23.03.2025).

81. Towards Safer Communities: Detecting Aggression and Offensive Language in Code-Mixed Tweets to Combat Cyberbullying / N. Nafis et al. *The 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, Canada. Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.woah-1.3 (дата звернення: 23.03.2025).

82. Balcıoğlu Y. S. Detecting Turkish Cyberbullying Tweets Using Machine Learning. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*. 2024. URL: https://doi.org/10.29130/dubited.1379657 (дата звернення: 23.03.2025).

83. Cyberbullying Detection: A Comparative Study of Classification Algorithms / P. Nuthalapati et al. *International Journal of Computer Science and Mobile Computing*. 2024. Vol. 13, no. 2. P. 1–12. URL: https://doi.org/10.47760/ijcsmc.2024.v13i02.001 (дата звернення: 23.03.2025).

84. Khairy M., Mahmoud T. M., Abd-El-Hafeez T. The effect of rebalancing techniques on the classification performance in cyberbullying datasets. *Neural Computing and Applications*. 2023. URL: https://doi.org/10.1007/s00521-023-09084-w (дата звернення: 23.03.2025).

85. Perera A., Fernando P. Cyberbullying Detection System on Social Media Using Supervised Machine Learning. *Procedia Computer Science*. 2024. Vol. 239. P. 506–516. URL: https://doi.org/10.1016/j.procs.2024.06.200 (дата звернення: 23.03.2025).

86. Deep Learning Algorithms with Adam Optimization for Detecting of Cyberbullying Comments. *Nanotechnology Perceptions*. 2024. Vol. 20, S3. URL: https://doi.org/10.62441/nano-ntp.v20is3.47 (дата звернення: 23.03.2025).

87. Cyberbullying Detection Using Deep Learning: A Comparative Study / M. Alkasassbeh et al. *2024 2nd International Conference on Cyber Resilience (ICCR)*, Dubai, United Arab Emirates, 26–28 February 2024. 2024. URL: https://doi.org/10.1109/iccr61006.2024.10533166 (дата звернення: 23.03.2025).

88. Albayari R., Abdallah S., Shaalan K. Cyberbullying Detection Model for Arabic Text Using Deep Learning. *Journal of Information & Knowledge Management*. 2024. URL: https://doi.org/10.1142/s0219649224500163 (дата звернення: 23.03.2025).

89. Nath S. S., Karim R., Miraz M. H. Deep Learning Based Cyberbullying Detection in Bangla Language. *Annals of Emerging Technologies in Computing*. 2024. Vol. 8, no. 1. P. 50–65. URL: https://doi.org/10.33166/aetic.2024.01.005 (дата звернення: 23.03.2025).

90. Potential Applications of Explainable Artificial Intelligence to Actuarial Problems / C. Lozano-Murcia et al. *Mathematics*. 2024. Vol. 12, no. 5. P. 635. URL: https://doi.org/10.3390/math12050635 (дата звернення: 23.03.2025).

91. Vilone G., Longo L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction*. 2021. Vol. 3, no. 3. P. 615–661. URL: https://doi.org/10.3390/make3030032 (дата звернення: 23.03.2025).

92. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence / S. Ali et al. *Information Fusion*. 2023. P. 101805. URL: https://doi.org/10.1016/j.inffus.2023.101805 (дата звернення: 23.03.2025).

93. Al-Ansari N., Al-Thani D., Al-Mansoori R. S. User- Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. *Human Behavior and Emerging Technologies*. 2024. Vol. 2024, no. 1. URL: https://doi.org/10.1155/2024/4628855 (дата звернення: 23.03.2025).

94. Gongane V. U., Munot M. V., Anuse A. Explainable AI for Reliable Detection of Cyberbullying. *2023 IEEE Pune Section International Conference (PuneCon)*, Pune, India, 14–16 December 2023. 2023. URL: https://doi.org/10.1109/punecon58714.2023.10450132 (дата звернення: 23.03.2025).

95. Aggarwal P., Mahajan R. Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification. *Journal of Information Systems and Informatics*. 2024. Vol. 6, no. 2. P. 607–623. URL: https://doi.org/10.51519/journalisi.v6i2.692 (дата звернення: 23.03.2025).

96. Garg P., Sharma M. K., Kumar P., Improving Hate Speech Classification Through Ensemble Learning and Explainable AI Techniques, *Arab. J. Sci. Eng.* (2024). URL: https://doi.org/10.1007/s13369-024-09540-2 (дата звернення: 23.03.2025).

97. Cyberbullying Sexism Harassment Identification by Metaheurustics-Tuned eXtreme Gradient Boosting / M. Dobrojevic et al. *Computers, Materials & Continua*. 2024. Vol. 80, no. 3. P. 4997–5027. URL: https://doi.org/10.32604/cmc.2024.054459 (дата звернення: 23.03.2025).

98. Aldhyani T. H. H., Al-Adhaileh M. H., Alsubari S. N. Cyberbullying Identification System Based Deep Learning Algorithms. *Electronics*. 2022. Vol. 11, no. 20. P. 3273. URL: https://doi.org/10.3390/electronics11203273 (дата звернення: 23.03.2025).

99. Kumar B. V. P., Sadanandam D. M. A Fusion Architecture of BERT and RoBERTa for Enhanced Performance of Sentiment Analysis of Social Media Platforms. *International Journal of Computing and Digital Systems*. 2024. Vol. 15, no. 1. P. 51–66. URL: https://doi.org/10.12785/ijcds/150105 (дата звернення: 23.03.2025).

100. 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. URL: https://doi.org/10.35546/kntu2078-4481.2024.4.45 (дата звернення: 19.03.2025).

101. Собко О. В. Прикладне застосування методу аналізу та формування репрезентативних вибірок текстових даних. *Інформаційні технології і автоматизація* : матеріали XVII Міжнар. наук.-практ. конф., м. Одеса, 31 жовт. – 1 листоп. 2024 р. / ОНТУ. Одеса, 2024. С. 687–689.

102. ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling / A. Alcoforado et al. *Lecture Notes in Computer Science*. Cham, 2022. P. 125–136. URL: https://doi.org/10.1007/978-3-030-98305-5_12 (дата звернення: 23.03.2025).

103. Собко О. В., Бармак О. В. Виявлення кіберзалякувань в інформаційному середовищі засобами машинного навчання. *Інформаційна, функційна і кібербезпека СКІФіК-2024 :* матеріали IV наук.-техн. конф., м. Харків, 29–30 листоп. 2024 р. Харків, 2024. С. 96–97.

104. Собко О. В. Підхід до нейромережевого виявлення та класифікації кіберзалякувань в освітньому процесі. *Сучасні інформаційні технології в освіті і науці* : зб. матеріалів VI Всеукр. наук.-практ. конф., м. Умань, 14–15 листоп. 2024 р. Умань, 2024. С. 219–222.

105. Garrido-Merchan E. C., Gozalo-Brizuela R., Gonzalez-Carvajal S., Comparing BERT Against Traditional Machine Learning Models in Text Classification, *J. Comput. Cogn. Eng.* 2.4 (2023) 352–356. URL: https://doi.org/10.48550/arXiv.2005.13012 (дата звернення: 23.03.2025).

106. Собко О. В. Виявлення та класифікація кіберзалякувань у цифрових текстах засобами штучного інтелекту. *Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах».* 2024. № 4.

С. 143–152. URL: https://doi.org/10.31891/2219-9365-2024-80-18 (дата звернення: 19.03.2025).

107. Собко О. В. Метод класифікації кіберзалякувань в україномовному текстовому контенті засобами штучного інтелекту. *Наука і техніка сьогодні*. 2024. № 13 (41). С. 1252–1263. URL: https://doi.org/10.52058/2786-6025-2024-13(41)-1252-1263 (дата звернення: 19.03.2025).

108. Sobko O. Practice Implementation of the Method for Analysis and Formation of Representative *Text Datasets. Information Technology and Implementation (Satellite)* : proceedings of the 11th Int. Conf., Kyiv, Ukraine, November 21, 2024. Kyiv, 2024. P. 88–89. (дата звернення: 19.03.2025).

109. Собко О. В. Метод аналізу та формування репрезентативних датасетів для виявлення кіберзалякувань у текстовому контенті. *Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій* : тези доп. XII Міжнар. наук.-практ. конф., Запоріжжя, 10–12 груд. 2024 р. / Нац. ун-т «Запоріз. політехніка». Запоріжжя, 2024. С. 402–406.

110. LSTM-SN: complex text classifying with LSTM fusion social network / W. Wei et al. *The Journal of Supercomputing*. 2023. URL: https://doi.org/10.1007/s11227-022-05034-w (дата звернення: 23.03.2025).

111. An efficient two-state GRU based on feature attention mechanism for sentiment analysis / M. Zulqarnain et al. *Multimedia Tools and Applications*. 2022. URL: https://doi.org/10.1007/s11042-022-13339-4 (дата звернення: 23.03.2025).

112. Classification of User's Review Using Modified Logistic Regression Technique / R. Reddy, U.M.A. Kumar. *International Journal of Systems Assurance Engineering and Managemen*t, 15, 2024, pp. 279–286. URL: https://doi.org/10.1007/s13198-022-01711-4 (дата звернення: 23.03.2025).

113. А. с. № 132726 Україна. Твір наукового характеру «Метод виявлення кібербулінгу у текстовому контенті» / О. В. Собко. 2025.

114. А. с. № 132727 Україна. Твір наукового характеру «Метод інтерпретації результатів нейромережевого виявлення кібербулінгу у текстовому контенті» / О. В. Собко. 2025.

115. Veziroğlu M., Eziroğlu E., Ömür Bucak İ. Performance comparison between naive bayes and machine learning algorithms for news classification. *Bayesian Inference – Recent Trends*. 2024. URL: https://doi.org/10.5772/intechopen.1002778 (дата звернення: 23.03.2025).

116. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network / O. Zalutska, M. Molchanova, O. Sobko, O. Mazurets, O. Pasichnyk, O. Barmak, I. Krak. *CEUR Workshop Proceedings*, 2023, vol. 3387, pp. 344–356. URL: https://ceur-ws.org/Vol-3387/paper26.pdf (дата звернення: 19.03.2025).

117. Sentiment Analysis using Support Vector Machine and Random Forest / T. Ahmed Khan et al. *Journal of Informatics and Web Engineering*. 2024. Vol. 3, no. 1. P. 67–75. URL: https://doi.org/10.33093/jiwe.2024.3.1.5 (дата звернення: 23.03.2025).

118. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification / V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak. *12th International Conference on Advanced Computer Information Technologies (ACIT 2022)*, 2022, pp. 400–405. URL: https://doi.org/10.1109/ACIT54803.2022.9913162 (дата звернення: 19.03.2025).

119. NLP- Based Speech Analysis Using K- Neighbor Classifier / A. Renuka, Bh. Rishu. 2024. URL: https://doi.org/10.1002/9781394175376.ch13 (дата звернення: 23.03.2025).

120. Wongvorachan T., He S., Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*. 2023. Vol. 14, no. 1. P. 54. URL: https://doi.org/10.3390/info14010054 (дата звернення: 23.03.2025).

121. EMFSA: Emoji-based multifeature fusion sentiment analysis / H. Tang et al. PLOS ONE. 2024. Vol. 19, no. 9. P. e0310715. URL: https://doi.org/10.1371/journal.pone.0310715 (дата звернення: 23.03.2025).

122. Собко О. В., Бармак О. В. Метод виявлення кіберзалякувань у текстовому контенті нейромережевими засобами. *Науковий журнал «Наукові праці Донецького національного технічного університету», серія «Проблеми моделювання та автоматизації проектування»*. 2025. № 1 (21). С. 52–61. URL: https://doi.org/10.31474/2074-7888 (дата звернення: 19.03.2025).

123. Собко О. В. Метод інтелектуального виявлення та класифікації кіберзалякувань у текстовому контенті. *Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024* : матеріали XII Міжнар. наук.-практ. конф., м. Одеса, 23–25 верес. 2024 р. Одеса, 2024. С. 262–265.

124. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers / M. Bayer et al. *International Journal of Machine Learning and Cybernetics*. 2022. URL: https://doi.org/10.1007/s13042-022-01553-3 (дата звернення: 23.03.2025).

125. Talaei Khoei T., Ould Slimane H., Kaabouch N. Deep learning: systematic review, models, challenges, and research directions. *Neural Computing and Applications*. 2023. URL: https://doi.org/10.1007/s00521-023-08957-4 (дата звернення: 23.03.2025).

126. Собко О. В. Інтерпретація результатів виявлення кіберзалякувань у текстах з використанням нейронних мереж. *Актуальні проблеми комп'ютерних наук АПКН-2024* : зб. наук. пр. за матеріалами XVI Всеукр. наук.-практ. конф., м. Хмельницький, 15–16 листоп. 2024 р. Хмельницький, 2024. С. 467–473.

127. Собко О. В. Візуальна інтерпретація нейромережевого виявлення кібербулінгу у цифрових текстах. *Нейромережні технології та їх застосування НМТі3-2024* : зб. наук. пр. XXIII Міжнар. наук. конф., Краматорськ-Тернопіль, 11–12 груд. 2024 р. / ДДМА. Краматорськ-Тернопіль, 2024. С. 138–144.

128. Model-Based Deep Learning / N. Shlezinger, J. Whang, Y. C. Eldar, A. G. Dimakis. *Proceedings of the IEEE*, 111(5), 2023, pp. 465–499. URL: https://doi.org/10.48550/arXiv.2012.08405 (дата звернення: 23.03.2025).

129. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP / M. Molchanova. In: *Scientific Achievements and Innovations as a Way to Success: Proc. XXI Int. Scientific and Practical Conf.*, May 1–3, 2024, Vilnius, Lithuania. Vilnius, 2024, pp. 73–77.

130. Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю / О. В. Мазурець. *Глушковські читання: матеріали XII Всеукр. наук.-практ. конф.*, Київ, 2023, с. 106–109.

131. Alissa S., Wald M. Text Simplification Using Transformer and BERT. *Computers, Materials & Continua*. 2023. Vol. 75, no. 2. P. 3479–3495. URL: https://doi.org/10.32604/cmc.2023.033647 (дата звернення: 23.03.2025).

132. Kiefer S. CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge. *Information Fusion*. 2022. Vol. 77. P. 184–195. URL: https://doi.org/10.1016/j.inffus.2021.07.014 (дата звернення: 23.03.2025).

133. A Confusion Matrix for Evaluating Feature Attribution Methods / A. Arias-Duart et al. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vancouver, BC, Canada, 17–24 June 2023. 2023. URL: https://doi.org/10.1109/cvprw59228.2023.00380 (дата звернення: 23.03.2025).

134. Confusion Matrices: A Unified Theory / J. Erbani et al. *IEEE Access*. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3507199 (дата звернення: 23.03.2025).

135. Relative Confusion Matrix: An Efficient Visualization for the Comparison of Classification Models / L. E. Pommé et al. *Artificial Intelligence and Visualization: Advancing Visual Knowledge Discovery*. Cham, 2024. P. 223–243. URL: https://doi.org/10.1007/978-3-031-46549-9_7 (дата звернення: 23.03.2025).

136. Rainio O., Teuho J., Klén R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*. 2024. Vol. 14, no. 1. URL: https://doi.org/10.1038/s41598-024-56706-x (дата звернення: 23.03.2025).

137. 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.

138. А. с. № 132042 Україна. Твір наукового характеру «Метод аналізу репрезентативності та формування навчальних вибірок» / О. В. Собко. 2024.

139. Tweet Files for Gender Guessing. *Kaggle*. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing (дата звернення: 23.03.2025).

140. Googletrans. *PyPI*. URL: https://pypi.org/project/googletrans (дата звернення: 23.03.2025).

141. TAG-it Dataset Distribution. *Live European Language Grid*. URL: https://live.european-language-grid.eu/catalogue/corpus/8112/download/ (дата звернення: 23.03.2025).

142. Cyberbullying Tweets. *Kaggle*. URL: https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets (дата звернення: 23.03.2025).

143. Національні демографічні прогнози. *Idss.org.ua*. URL: https://idss.org.ua/forecasts/nation_pop_proj (дата звернення: 23.03.2025).

144. Демографічний прогноз по Україні 2023. *Idss.org.ua*. URL: https://idss.org.ua/arhiv/Ukr1991s1%202023%202035.xls (дата звернення: 23.03.2025).

145. А. с. № 132918 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення кіберзалякувань у текстовому контенті» / О. В. Собко. 2025.

146. Benchmarking Language Models for Cyberbullying Identification and Classification from Social-Media Texts / K. Verma, T. Milosevic, K. Cortis, B. Davis. *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, 2022, pp. 26–31. URL: https://aclanthology.org/2022.lateraisse-1.4/ (дата звернення: 23.03.2025).

147. 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: 23.03.2025).

148. We Need to Talk About Classification Evaluation Metrics in NLP / P. Vickers et al. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Nusa Dua, Bali. Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.ijcnlp-main.33 (дата звернення: 23.03.2025).

149. Deep Learning for Multi-Labeled Cyberbully Detection: Enhancing Online Safety / N. Islam et al. *2023 International Conference on Data Science and Network Security (ICDSNS)*, Tiptur, India, 28–29 July 2023. 2023. URL: https://doi.org/10.1109/icdsns58469.2023.10245135 (дата звернення: 23.03.2025).

150. Enhancing Online Safety: Natural Language Processing Based Multi-Label Cyberbullying Classification in Bangla / M. Saifuddin et al. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 13–15 December 2023. 2023. URL: https://doi.org/10.1109/iccit60459.2023.10441466 (дата звернення: 23.03.2025).

151. Marchenko O., Isoieva M. Automatic Generation of Coherent Natural Language Texts. *Flexible Query Answering Systems*. Cham, 2023. P. 79–92. URL: https://doi.org/10.1007/978-3-031-42935-4_7 (дата звернення: 23.03.2025).

152. Skurzhanskyi O. H., Marchenko O. O., Anisimov A. V. Specialized Pre-Training of Neural Networks on Synthetic Data for Improving Paraphrase Generation. *Cybernetics and Systems Analysis*. 2024. URL: https://doi.org/10.1007/s10559-024-00658-7 (дата звернення: 23.03.2025).

153. Information Technology for Adaptive Semantic Testing of Knowledge Level of Educational Materials / O. Mazurets et al. *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Cracow, Poland, 22–25 September 2021. 2021. URL: https://doi.org/10.1109/idaacs53288.2021.9661043 (дата звернення: 23.03.2025).

154. Krak I., Barmak O., Manziuk E. Using visual analytics to develop human and machine- centric models: A review of approaches and proposed information technology. *Computational Intelligence*. 2020. URL: https://doi.org/10.1111/coin.12289 (дата звернення: 23.03.2025).

155. Ontology-Driven Lexicographic Systems / M. Nadutenko et al. *Lecture Notes in Networks and Systems*. Cham, 2022. P. 204–215. URL: https://doi.org/10.1007/978-3-030-98012-2_16 (дата звернення: 02.04.2025).

156. Transdisciplinary Principles of Narrative Discourse as a Basis for the Use of Big Data Communicative Properties / O. Stryzhak et al. *Advances in Intelligent Systems and Computing*. Cham, 2021. P. 258–273. URL: https://doi.org/10.1007/978-3-030-73103-8_17 (дата звернення: 02.04.2025).

157. Decision-making System Based on The Ontology of The Choice Problem / O. Stryzhak et al. *Journal of Physics: Conference Series*. 2021. Vol. 1828, no. 1. P. 012007. URL: https://doi.org/10.1088/1742-6596/1828/1/012007 (дата звернення: 02.04.2025).

158. Intelligent Method for Classifying the Level of Anthropogenic Disasters / K. Lipianina-Honcharenko et al. *Big Data and Cognitive Computing*. 2023. Vol. 7, no. 3. P. 157. URL: https://doi.org/10.3390/bdcc7030157 (дата звернення: 02.04.2025).

159. Evaluation of the Effectiveness of Machine Learning Methods for Detecting Disinformation in Ukrainian Text Data. Kh. Lipianina-Honcharenko, M. Soia, Kh. Yurkiv, A. Ivasechko. *Proceedings of the Seventh International Workshop on Computer Modeling and Intelligent Systems (CMIS–2024)*. 2024. P. 97–109. URL: https://ceur-ws.org/Vol-3702/paper9.pdf (дата звернення: 02.04.2025).

160. OLTW-TEC: online learning with sliding windows for text classifier ensembles / K. Lipianina-Honcharenko et al. *Frontiers in Artificial Intelligence*. 2024. Vol. 7. URL: https://doi.org/10.3389/frai.2024.1401126 (дата звернення: 13.04.2025).

161. Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques / C. Raj et al. *Electronics*. 2021. Vol. 10, no. 22. P. 2810. URL: https://doi.org/10.3390/electronics10222810 (дата звернення: 23.03.2025).

162. 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 (дата звернення: 23.03.2025).

163. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages / A. ImaniGooghari et al. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Toronto, Canada. Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.acl-long.61 (дата звернення: 23.03.2025).

ДОДАТОК А.
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

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.

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. Tymofiiev, 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.

ДОДАТОК Б.
ДОВІДКИ ТА АКТ ПРО ВПРОВАДЖЕННЯ

ДОВІДКА

про впровадження результатів дисертаційної роботи
Собко Олени Віталіївни
«Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»

В діяльності відділу протидії кіберзлочинам в Хмельницькій області Департаменту кіберполіції Національної поліції України знайшли застосування наступні результати дисертаційної роботи здобувачки наукового ступеня доктора філософії Собко Олени Віталіївни «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»:

– метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості;

– метод виявлення кіберзалякувань у текстовому контенті;

– метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті.

У відділі протидії кіберзлочинам в Хмельницькій області зазначені результати були використані при аналізі вияалення в підозрілих текстових матеріалах вебджерел фактів наявності кіберзалякувань із визначенням типів кіберзалякувань та візуальним поясненням прийнятих рішень щодо виявлених фактів кіберзалякувань. Тестування й практичне використання запропонованих методів показало високу точність виявлення агресивних і образливих соціально неприйнятних проявів залякувань у вебпросторі; підтвердило наочність візуальних пояснень результатів аналізу текстових матеріалів для детектування агресії, образ та кібербулінгу.

Начальник відділу
протидії кіберзлочинам
в Хмельницькій області
Департаменту кіберполіції
Національної поліції України _____ Дядик О.М.
26.03.2025

ПП «Авіві» ЄДРПОУ 35605257, UA413510050000026000065710440
www.avivi.ua hello@avivi.pro +380 68 2020200
м. Хмельницький, вул. Подільська, 109.

ДОВІДКА

про впровадження результатів дисертаційної роботи
Собко Олени Віталіївни
«Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»

Наступні результати дисертаційної роботи Собко Олени Віталіївни «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» були впроваджені у виробничу діяльність ПП «Авіві»: метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості; метод виявлення кіберзалякувань у текстовому контенті; метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті.

Зазначені результати були використані при розробці соціально орієнтованих сервісів, для автоматизованого виявлення кіберзалякувань у текстових повідомленнях, з деталізацією у вигляді визначення типів кіберзалякувань та візуальним поясненням прийнятих рішень щодо виявлених типів кіберзалякувань. Додатково в межах підготовки навчальних даних для нейромережі було застосовано метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечило збалансованість навчальних даних для автоматизованого виявлення кіберзалякувань. Тестування й практична експлуатація створеного програмного забезпечення підтвердили високу точність виявлення та класифікації кіберзалякувань, а також наочність візуальної інтерпретації одержаних рішень щодо наявності образливих або агресивних включень.

Директор ПП «Авіві» Аскеров В.В. 20.01.2025

29001
м.Хмельницький,
вул. Подільська
буд.109

+380 (63) 397 55 35
itkmuacluster@gmail.com
www.it.km.ua

громадська організація
“IT-кластер
міста
Хмельницького”

ДОВІДКА
про впровадження результатів дисертаційної роботи
Собко Олени Віталіївни «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»

Результати дисертаційної роботи Собко Олени Віталіївни на тему «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» були використані під час виробничої діяльності громадської організації «IT-кластер міста Хмельницького» при розробці компонентів системи обміну корпоративними повідомленнями. Зокрема, були впроваджені такі результати з дисертаційної роботи: метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечує недискримінацію за віковою, соціальною, гендерною, релігійною приналежністю, що дозволив забезпечити релевантний підбір датасетів для виявлення кіберзалякувань; метод виявлення кіберзалякувань у текстовому контенті, що дав можливість підвищити якість виявлення кіберзалякувань та класифікувати за їх; метод інтерпретації результатів виявлення кіберзалякувань, що дозволив подавати результати в зрозумілому для користувача вигляді.

Розроблене на основі зазначених результатів програмне забезпечення забезпечило автоматизоване виявлення проявів кіберзалякуваннь та їх типізацію. Зокрема, програмна система при аналізі текстових повідомлень визначає агресивні висловлювання та потенційно токсичні комунікаційні патерни, а також надає обґрунтоване пояснення висновків шляхом візуальної інтерпретації результатів. За результатом впровадження в систему обміну корпоративними повідомленнями громадської організації «IT-кластер міста Хмельницького», наведене дозволило ефективно моніторити та таргетувати ризики агресивної взаємодії між працівниками, сприяючи створенню більш толерантного та етичного робочого середовища, що позитивним чином вплинуло на продуктивність робочого процесу.

Заступник голови ГО
«IT-кластер міста Хмельницького» *Сергій ЯЦИШЕН* *25.01.2025*

ДОВІДКА

про впровадження результатів дисертаційної роботи
Собко Олени Віталіївни «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»

Результати дисертаційної роботи здобувачки наукового ступеня доктора філософії Собко Олени Віталіївни на тему «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» були впроваджені в діяльність товариства з обмеженою відповідальністю «Системи для бізнесу 2». При розробці програмного забезпечення товариством з обмеженою відповідальністю «Системи для бізнесу 2» було використано розроблені метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, метод виявлення кіберзалякувань у текстовому контенті, метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті.

Товариство з обмеженою відповідальністю «Системи для бізнесу 2» інтегрувало розроблені методи здобувачки наукового ступеня доктора філософії Собко Олени Віталіївни у корпоративні системи обміну повідомленнями, впроваджуючи механізми автоматичного аналізу текстових повідомлень на предмет агресивної та деструктивної поведінки. Запропоновані методи дозволили створити систему аналізу текстового контенту системи обміну повідомленнями на кіберзалякування, яка не лише виявляє, класифікує та фільтрує потенційно шкідливий контент, а й за потреби надає детальне пояснення щодо виявлених випадків кіберзалякувань шляхом візуальної інтерпретації. Наведене значно підвищило рівень безпеки комунікації між співробітниками підприємства, сприяючи формуванню здорового та етичного робочого середовища.

Директор ТОВ «Системи для бізнесу 2» Третько С.В. 10.01.2025

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Хмельницького
національного університету
д.т.н., професор Олег СИНЮК

"27" листопада 2024 р.

АКТ

про впровадження в навчальний процес результатів досліджень
аспірантки Собко Олени Віталіївни за темою дослідження
«Методи виявлення та класифікації кіберзалякувань у текстовому контенті
засобами штучного інтелекту»

Результати дисертаційної роботи Олени Собко, а саме, методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту, використовуються в навчальному процесі при викладанні дисциплін бакалаврського рівня «Методи та системи штучного інтелекту», «Інтелектуальний аналіз даних» та магістерського рівня «Моделі та методи інтелектуального аналізу текстової інформації та машинного навчання», виконанні кваліфікаційних робіт здобувачів бакалаврського та магістерського рівнів вищої освіти спеціальності "Комп'ютерні науки".

Акт обговорений і схвалений на засіданні кафедри комп'ютерних наук (протокол № 4 від 27 листопада 2024 р.).

Зав. каф. комп'ютерних наук,
д.т.н., професор Олександр БАРМАК

Секретар каф. комп'ютерних наук,
ст.викладач Тетяна СКРИПНИК

ДОДАТОК В.
АВТОРСЬКІ СВІДОЦТВА

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 132920

Комп'ютерна програма «Інтелектуальна інформаційна система для оцінювання та коригування репрезентативності текстових датасетів»

(вид, назва твору)

Автор (співавтори) **Собко Олена Віталіївна**

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать повністю **Собко Олена Віталіївна, [redacted], [redacted], м. Хмельницький, 29019**

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

Дата реєстрації 3 лютого 2025 р.

Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій» **Олена ОРЛЮК**

УКРНОІВІ

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 132921

Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»

(вид, назва твору)

Автор (співавтори) **Собко Олена Віталіївна**

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать повністю **Собко Олена Віталіївна, [redacted], [redacted], м. Хмельницький, 29019**

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

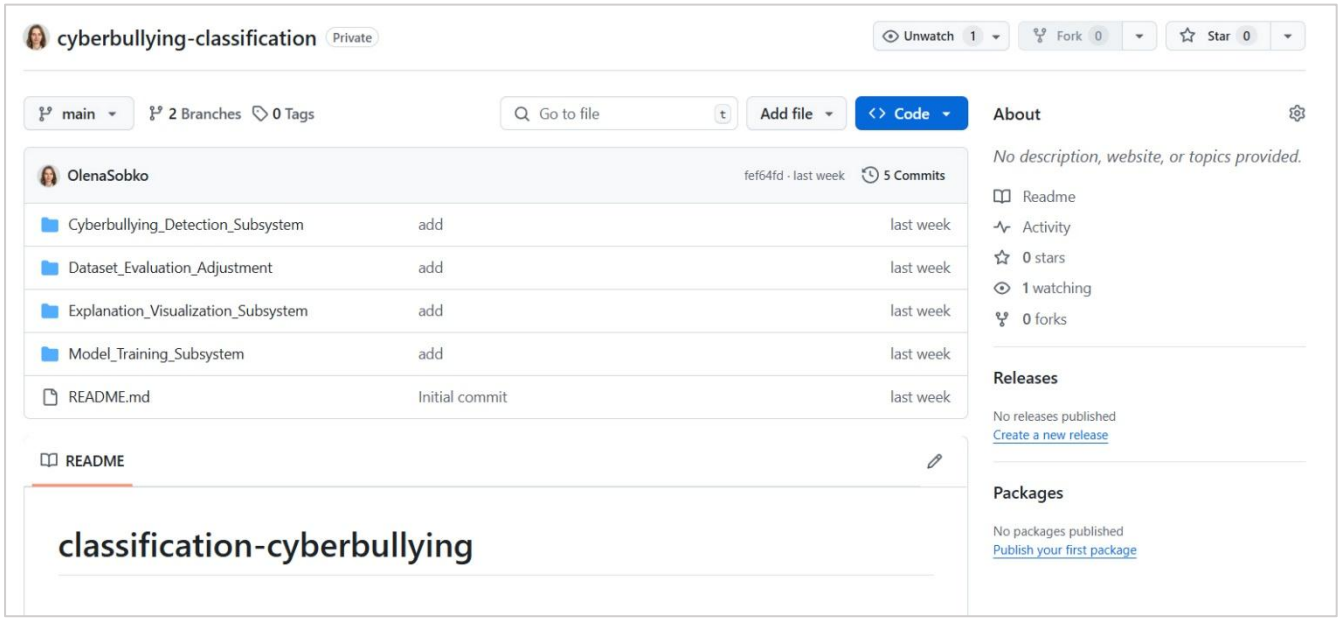
Дата реєстрації 3 лютого 2025 р.

Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій» **Олена ОРЛЮК**

УКРНОІВІ

ДОДАТОК Г.
ПРОГРАМНИЙ КОД

Програмний код, що створений під час роботи над дисертаційним дослідженням, доступний у репозиторії GitHub: https://github.com/OlenaSobko/cyberbullying-classification (дата звернення: 02.04.2025). На рисунку наведено світлину екрана репозиторію “cyberbullying-classification”.



“Dataset Evaluation and Adjustment Subsystem” – містить коди для оцінювання та балансування датасету, включаючи “DataPreprocessor”, “DatasetFairnessEvaluator» та «DatasetBalancer”.

“Cyberbullying Detection Subsystem” – включає коди для виявлення кіберзалякувань, зокрема “TextPreprocessor” та “PredictionEngine”, який використовує моделі BiLSTM і BERT.

“Explanation and Visualization Subsystem” – містить коди для пояснення рішень моделі за допомогою LIME “ExplanationGenerator” та візуалізації результатів “VisualizationEngine”.

“Model Training Subsystem” – включає код для навчання та валідації моделей, зокрема “DataProcessor”, “ModelTrainer”, “HyperparameterOptimizer” і “EvaluationEngine”.