DOI 10.31891/2307-5732-2024-331-17
УДК 004.8

МОЛЧАНОВА МАРИНА
Хмельницький національний університет
https://orcid.org/0000-0001-9810-936X
e-mail: m.o.molchanova@gmail.com
МАЗУРЕЦЬ ОЛЕКСАНДР
Хмельницький національний університет
https://orcid.org/0000-0002-8900-0650
e-mail: exe.chong@gmail.com
СОБКО ОЛЕНА
Хмельницький національний університет
https://orcid.org/0000-0001-5371-5788
e-mail: olena.sobko.ua@gmail.com
ВІТ РОМАН
Хмельницький національний університет
https://orcid.org/0009-0009-6958-4730
e-mail: vit.roman.vit@gmail.com
НАЗАРОВ В'ЯЧЕСЛАВ
Ліцей №17 Хмельницької міської ради
e-mail: uuu.vich228@gmail.com

АЛГОРИТМ ВИЯВЛЕННЯ АБ'ЮЗИВНОГО ВМІСТУ В УКРАЇНОМОВНОМУ АУДІОКОНТЕНТІ ДЛЯ ІМПЛЕМЕНТАЦІЇ В ОБ'ЄКТНО-ОРІЄНТОВАНУ ІНФОРМАЦІЙНУ СИСТЕМУ

У роботі пропонується практичний підхід до виявлення аб'юзивного вмісту в україномовному аудіоконтенті на основі нового алгоритму, що використовує статистичний та нейромережевий підходи для виявлення аб'юзивного вмісту. Наведено результати аналізу практичної ефективності запропонованого підходу, що підтверджують спроможність виявлення аб'юзивного вмісту в україномовному аудіоконтенті.

Ключові слова: виявлення аб'юзивного вмісту, RNN, аудіоконтент, об'єктно-орієнтований підхід.

MOLCHANOVA MARYNA, MAZURETS OLEKSANDR, SOBKO OLENA, VIT ROMAN
Khmelnytskyi National University
NAZAROV VIACHESLAV
Khmelnytskyi Lyceum №17

ALGORITHM FOR DETECTION OF ABUSIVE CONTENT IN AUDIO CONTENT FOR IMPLEMENTATION IN OBJECT-ORIENTED INFORMATION SYSTEM

The paper proposes the basic principles of developing an object-oriented information system for detecting abusive content in Ukrainian-language audio content based on a new algorithm that uses statistical and neural network approaches to detect abusive content. Detecting offensive content in text and audio content is an urgent task, as it helps to create a safe and healthy environment for communication, especially in online platforms. Offensive content can harm the people who hear or read it and violate their rights. It can also have a negative impact on society, contributing to the spread of hatred and violence.

To detect abusive speech in audio content, the proposed approach uses two key components: the use of dictionary methods and the analysis of the emotional tonality of utterances. A set of reviews was used as a dataset to determine the abusive component of the content, which was expanded by the authors by adding words of abuse.

An object-oriented information system architecture written in the Python programming language in the PyCharm programming environment is proposed. The information system consists of a software module for training recurrent neural network models and further saving trained instances, and a software module for detecting abusive content in Ukrainian-language audio content using trained RNN models. Since the recurrent neural network is trained on a short text data set, the system is less efficient at identifying texts that have a larger number of words.

In the example of the proposed approach, the accuracy of detecting offensive content is more than 90%. This means that the algorithm works correctly in the absence of offending highlights in the test data set. The results of the analysis of the effectiveness of the proposed approach show that in the vast majority of cases the conclusions regarding the acceptability of audio content based on the level of abuse are correct.

Keywords: abusive content detection, RNN, audio content, object-oriented approach.

Аналіз предметної області

Кількість користувачів соціальних мереж продовжує невпинно зростати, що свідчить про значний обсяг аудіоконтенту та відеоконтенту, що був опублікований і публікується щодня в подібних сервісах та соціальних площадках. У зв'язку з цим зростає і важливість забезпечення якості цього аудіоконтенту. Зокрема, важливим аспектом є своєчасне виявлення аб'юзивних проявів, оскільки несанкціонована, образлива або неправдива інформація може завдати шкоду та нанести негативні наслідки для людей, груп чи суспільства загалом.

Виявлення аб'юзивного контенту в текстовому та аудіоконтенті є актуальною задачею, адже це не лише сприяє створенню безпечного та здорового середовища для спілкування, особливо в онлайн-платформах, а й сприяє боротьбі з цькуванням та дискримінацією, оскільки аудіоповідомлення можуть містити образливу мову та здатні заподіяти шкоду слухачам [1]. Також це дозволяє оперативно реагувати на неприпустиму поведінку оточуючих, підвищуючи якість комунікацій і знижуючи ризик поширення

токсичного контенту. Виявлення аб'юзу в аудіоконтенті має важливе значення для дотримання норм законодавства та етики, оскільки деякі його види можуть бути не просто неприємними, а й незаконними.

Аб'юзивний контент вважається будь-який тип контенту, який містить образливі, шкідливі, неприйнятні або агресивні елементи, що можуть завдати шкоди іншим особам. В науковому контексті, аб'юзивний контент часто аналізується через призму його впливу на індивідів та суспільство, а також через механізми його поширення та виявлення [2].

Для виявлення аб'юзивного мовлення у аудіоконтенті підхід буде зосереджено на двох ключових елементах: використанні словникових методів та аналізі емоційної тональності висловлювань.

Останні публікації

Проблемою виявлення образливого вмісту наразі займається велика кількість науковців по всьому світу. Європейський суд з прав людини (CEDH), дає визначення мови ненависті, як «усі форми вираження, усні чи письмові, які поширюють, підбурюють, сприяють чи виправдовують ненависть на ґрунті нетерпимості» [3]. У роботі [4] наведено аналіз специфіки спілкування інтернет-користувачів. Досліджено важливі аспекти спілкування в соціальних мережах, зокрема, проблему виявлення образливого вмісту в повідомленнях користувачів. Проаналізовано характеристики спілкування в цьому контексті та визначено чинники, які можуть ускладнювати автоматизовану детекцію образливого контенту. Розглянуті етапи обробки природномовних текстових даних, зокрема, їхнє визначення як образливого. Досліджено можливі модифікації для врахування особливостей спілкування в соціальних мережах та інших факторів, які можуть впливати на точність класифікації текстових даних. Виявлено, що особливості повідомлень та контексту в соціальних мережах, такі як символи, цифри, емодзі та взаємозв'язки між користувачами, можуть бути враховані при автоматизованому виявленні образливого вмісту. Запропоновано метод на основі машинного навчання з модифікованим підходом до оброблення текстових даних та визначено формат вхідних даних для ефективного виявлення образливого вмісту в текстових повідомленнях у соціальних мережах.

У дослідженні [5] авторами розробилено лексикон ненависті мовою урду, на основі якого сформулювано анотований набір даних із 10 526 твітів на урду. Крім того, як базові експерименти авторами використано різні методи машинного навчання для виявлення ненависті. Використано перехідне навчання, щоб застосовувати для завдання попередньо підготовлені вбудовування слів FastText урду та багатомовні вбудовування BERT. Нарешті, виконано експеримент з чотирма різними варіантами BERT, з перехідним навчанням, і показано, що BERT, xlm-roberta та distil-Bert здатні досягти заохочувальних оцінок F1 0,68, 0,68 та 0,69 відповідно. Усі ці моделі різною мірою показали успіх, а також перевершили низку базових моделей глибокого та машинного навчання.

Метою роботи є розробка об'єктно-орієтованої інформаційної системи виявлення аб'юзивного вмісту в україномовному аудіоконтенті на основі запропонованого алгоритму, що використовує статистичний та нейромережевий підходи для виявлення аб'юзивного вмісту.

Основна частина

Алгоритм виявленні аб'юзивного вмісту в аудіоконтенті Для виявлення аб'юзивного вмісту в україномовному текстовому та аудіоконтенті слід визначити кожну складову, що вказує на наявність таких проявів. Як вже було зазначено, аб'юзія ефективно виявляється за наявністю в тексті таких ознак:

- аб'юзивне мовлення (використання слів аб'юзії);
- негативна емоційна тональність.

В загальному, алгоритм виявлення аб'юзивних проявів у аудіоконтенті наведено на Рис. 1.

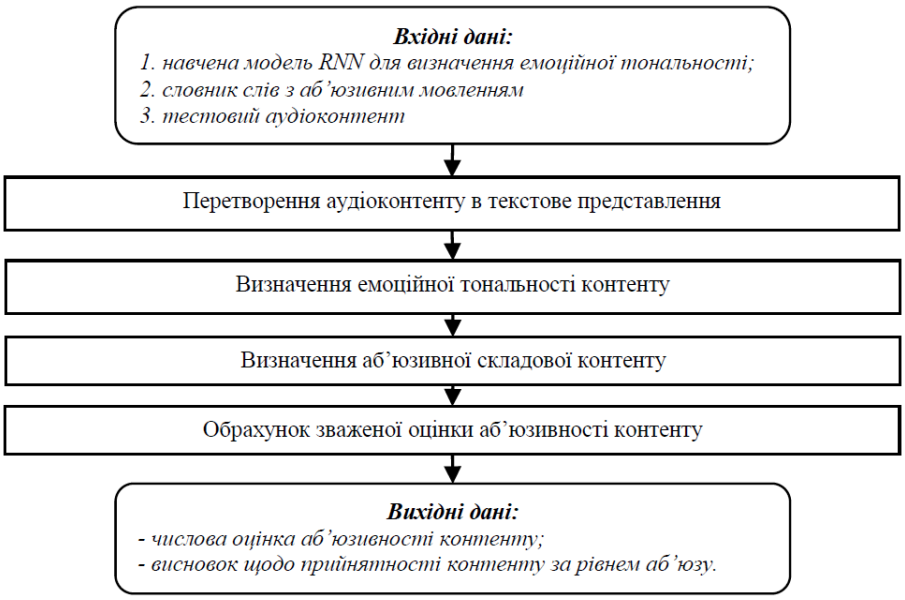

Рис. 1. Алгоритм виявленні аб'юзивного вмісту в аудіоконтенті

Вхідними даними для визначенні аб'юзивного вмісту у аудіоконтенті є попередньо навчена модель RNN для визначення емоційної тональності, словник слів з аб'юзивним мовленням та тестовий аудіоконтент.

Першим кроком є перетворення аудіоконтенту у текстове представлення за допомогою технології розпізнавання мови, що є процесом перетворення мовленнєвого сигналу в текстовий потік [6].

Наступним кроком є визначення емоційної тональності перетвореного аудіоконтенту у текстовий. Для цього використовується попередньо навчена модель RNN. Емоційна тональність визначається у проміжку від 0 до 1, де 0 – емоційно-негативна тональність, а 1 – емоційно-позитивна тональність.

Визначення аб'юзивної складової контенту відбувається статистичним шляхом та розраховується за формулою:

$$AbusiveSpeach = \frac{Abusive_{word}}{Len(Text)},$$

де $Abusive_{word}$ – кількість слів аб'юзії, що використано у текстовому представленні аудіоконтенту, що містяться у словнику слів аб'юзії, *Len(Text)* – загальна кількість слів у перетвореному аудіоконтенті у текстовий формат.

Обрахунок зваженої оцінки аб'юзивн 1. Mann S., Arora J., Bhatia M. Twitter Sentiment Analysis Using Enhanced BERT. Intelligent Systems and Applications. Springer. 2023. № 959. R. 263–271.

2. Slobodzian V., Molchanova M., Kovalchuk O. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. 12th International Conference on Advanced Computer Information Technologies. ACIT. 2022. R. 400–405.

3. European Court of Human Rights. Knowledge Sharing. 2023. https://www.echr.coe.int/Pages/home.aspx?p=caselaw/otherpublications&c=#n 15930944601351434310567_pointer.

4. Zabolotnia T. M., Sokolovska A. V. Metod avtomatyzovanoho vyznachennia naiavnosti obrazlyvoho vmistu tekstovykh povidomlen u sotsialnykh merezhakh. Visnyk ZhDTU. Seriia "Tekhnichni nauky". 2018. № 81. S. 103–108.

5. Ali R., Farooq U., Arshad U. Hate speech detection on Twitter using transfer learning. Computer Speech & Language. 2022. https://www.sciencedirect.com/science/article/abs/pii/S0885230822000110.

6. Zalutska O., Molchanova M., Sobko O. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings. 2023. № 3387. R. 561–571.

7. Abusive Language Dataset. 2023. https://hatespeechdata.com/#Ukranian-header.ості контенту відбувається з урахуванням нейромережевої оцінки емоційної тональності та аб'юзивної складової контенту та обраховується за формулою:

$$Abusive = \frac{k_1 \cdot AbusiveSpeach + k_2 \cdot (1 - SentRNN)}{2},$$

де *AbusiveSpeach* – статистична аб'юзивна складова контенту, *SentRNN* – нейромережева оцінка емоційної тональності перетвореного аудіоконтенту. k_1, k_2 – вагові коефіцієнти значимості показників. Обираються експериментальним шляхом, проте $k_1 + k_2 = 2$.

Вихідними даними є числова оцінка аб'юзивності контенту та висновок щодо прийнятності контенту за рівнем аб'юзу. Висновок прийнятності контенту за рівнем аб'юзу формується за гранично-заданим експертом значенням.

Таким чином, запропонований алгоритм виявлення аб'юзивних проявів в україномовному аудіоконтенті дозволяє одержувати числову оцінку аб'юзивності контенту та висновок його прийнятності за рівнем наявного аб'юзу.

Дані дослідження Для навчання нейромережі RNN, що буде в подальшому використана для ідентифікації емоційної тональності контенту, у якості експериментальних даних було використано набір даних відгуків з платформи «hotline». Загалом датасет складається із 7656 документів, де в навчальній вибірці знаходиться 6655 документів, і з них 1331 документ використано для валідації (що складає 20 % навчальної вибірки). Особливістю дата сету є те, що в ньому зустрічаються русизми, слова-покручі [6], тексти, максимально близькі до природної української мови.

Для визначення аб'юзивної складової контенту було використано набір даних «AbusiveLanguageDataset» [7]. Набір даних містить набір слів аб'юзу з нецензурною лексикою. Словник для української мови у базовому варіанті склав 623 слова, проте у ході дослідження був розширений авторами до 956 слів.

Об'єктно-орієнтована архітектура Інформаційна система розроблена з використанням мови програмування Python та середовища програмування PyCharm. Інформаційна система складається з програмного модуля для навчання моделей рекурентних нейромереж і подальшого збереження навчених екземплярів, та програмного модуля для виявлення аб'юзивного вмісту в україномовному аудіоконтенті з використанням навчених моделей RNN.

Об'єктно-орієнтована архітектура модуля для навчання моделей рекурентних нейромереж наведена на рисунку 2, а архітектура модуля для виявлення аб'юзивного вмісту в україномовному аудіоконтенті наведена на рисунку 3.

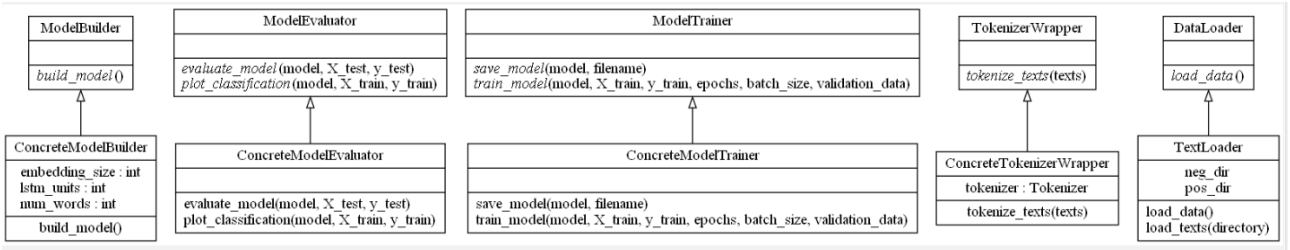

Рис. 2. Об'єктно-орієнтована архітектура модуля для навчання моделей RNN

Архітектура модуля для навчання моделей RNN використовує абстрактні класи та реалізації, щоб забезпечити більшу гнучкість, інкапсуляцію та можливість зміни компонентів системи. Кожен клас є компонентом, який може бути легко замінений або розширений.

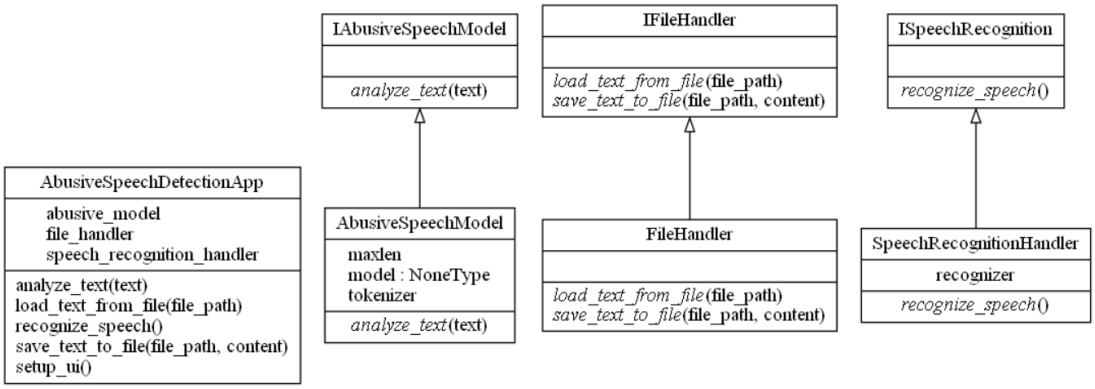

Рис. 3. Об'єктно-орієнтована архітектура модуля виявлення аб'юзивного вмісту

Структура модуля виявлення аб'юзивного вмісту розділяє обов'язки моделі, розпізнавання мовлення та обробки файлів на окремі класи, дотримуючись принципу єдиної відповідальності. Залежності впроваджуються в основний клас програми відповідно до принципу інверсії залежностей. Інтерфейси допомагають визначити чіткі контракти для кожного компонента, сприяючи інкапсуляції.

Дослідження ефективності Дослідження ефективності інформаційної системи виявлення аб'юзивного вмісту виконано за результатами застосування розроблених програмних модулів (Рис. 4).

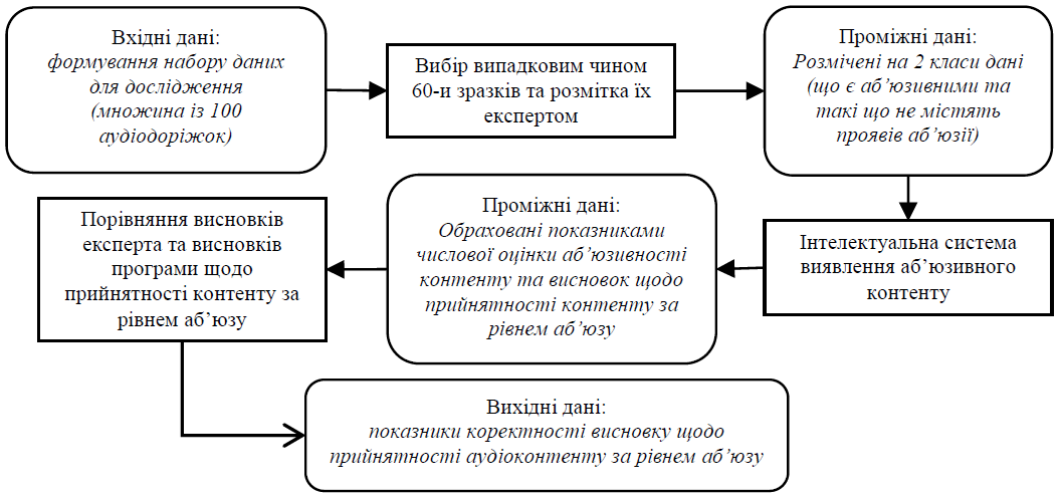

Рис. 4. Схема проведення експерименту з дослідження ефективності інформаційної системи

Для експерименту, з використанням інформаційної системи виявлення аб'юзивного вмісту було додатково зібрано 100 аудіотреків, тривалістю від 10 до 60 секунд.

Далі здійснюється вибір 60-и зразків, що обираються випадковим чином та проходять розмітку експертом. У якості експерта було задіяно як людину, так і програмний засіб ChatGPT 3.5 (було отримано 2 альтернативно-розмічених набори для подальшої класифікації програмою).

Після чого розмічені набори даних подавались інформаційної системи виявлення аб'юзивного вмісту з метою отримання обрахованої числової оцінки аб'юзивності аудіоконтенту та висновок щодо його прийнятності за рівнем аб'юзу.

Наступним етапом експерименту відбувається перевірка коректності висновків програмної реалізації відносно висновків експерта щодо прийнятності контенту за рівнем аб'юзу. Вихідними даними експерименту є показники коректності висновку щодо прийнятності аудіоконтенту за рівнем аб'юзу.

Результати експерименту Результати проведеного експерименту з дослідження ефективності інформаційної системи виявлення аб'юзивного вмісту на основі запропонованого алгоритму наведено у таблиці 1. Наведені результати свідчать, що в переважній більшості випадків висновки щодо прийнятності аудіоконтенту за рівнем аб'юзу коректні.

Таблиця 1

Кількісні результати коректності висновків щодо прийнятності аудіоконтенту

Тип результату	Результати за коректністю висновків		
	Експерт людина	Експерт ChatGPT	Всього
Коректно	56	53	109
Не коректно	4	7	11
Загалом	60	60	120

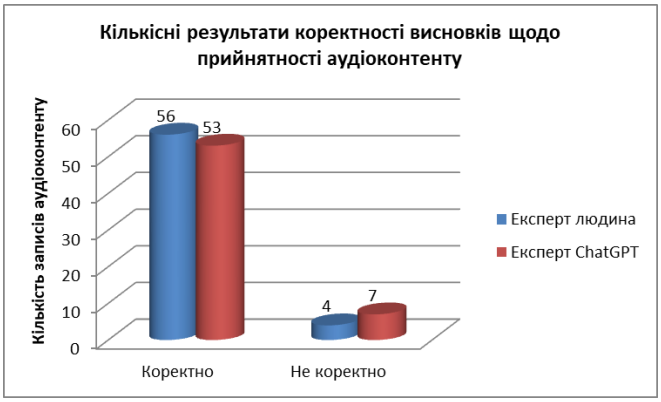

Рис. 5. Діаграма результатів висновків щодо прийнятності аудіоконтенту за рівнем аб'юзу

Розроблена інформаційна система на основі запропонованого алгоритму має обмеження. Оскільки рекурентна нейронна мережа навчена на наборі текстових даних короткої послідовності (до 200 слів, середня довжина тестових прикладів 17 слів), система менш ефективно ідентифікує тексти, які мають довжину понад 200 слів. Для підвищення ефективності у таких випадках необхідно розширити навчальні набори більш довгими текстовими даними та перенавчити нейронну мережу.

Висновки

У роботі було розглянуто сучасний стан напряму виявлення аб'юзивного мовлення у аудіоконтенті. На основі проведеного аналізу предметної області, запропоновано алгоритм виявлення аб'юзивних проявів у аудіоконтенті. Для виявлення аб'юзивного мовлення у аудіоконтенті у запропонованому підході буде використано два ключових елемента: використання словникових методів та аналізу емоційної тональності висловлювань.

У якості даних дослідження було використано набір відгуків з платформи «hotline», розміром 7656 документів, що призначений для навчання нейромережі RNN для ідентифікації емоційної тональності контенту та набір даних «AbusiveLanguageDataset», що був розширений авторами до 956 слів аб'юзії та використовується для визначення аб'юзивної складової контенту.

Запропоновано об'єктно-орієнтовану архітектуру інформаційної системи, що написана мовою програмування Python у середовищі програмування PyCharm. Інформаційна система складається з програмного модуля для навчання моделей рекурентних нейромереж і подальшого збереження навчених екземплярів, та програмного модуля для виявлення аб'юзивного вмісту в україномовному аудіоконтенті з використанням навчених моделей RNN.

Запропонований підхід має деякі обмеження. Оскільки рекурентна нейронна мережа навчена на наборі текстових даних короткої послідовності (до 200 слів, середня довжина тестових прикладів 17 слів), система менш ефективно ідентифікує тексти, які мають довжину понад 200 слів. Для підвищення ефективності у таких випадках необхідно розширити навчальні набори більш довгими текстовими даними та перенавчити нейронну мережу, на що будуть спрямовані подальші дослідження.

Література

1. Mann S., Arora J., Bhatia M. Twitter Sentiment Analysis Using Enhanced BERT. Intelligent Systems and Applications. Springer. 2023. № 959. P. 263–271.

2. Slobodzian V., Molchanova M., Kovalchuk O. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. 12th International Conference on Advanced Computer Information Technologies. ACIT. 2022. P. 400–405.

3. European Court of Human Rights. Knowledge Sharing. 2023. https://www.echr.coe.int/Pages/home.aspx?p=caselaw/otherpublications&c=#n 15930944601351434310567_pointer.

4. Заболотня Т. М., Соколовська А. В. Метод автоматизованого визначення наявності образливого вмісту текстових повідомлень у соціальних мережах. Вісник ЖДТУ. Серія "Технічні науки". 2018. № 81. С. 103–108.

5. Ali R., Farooq U., Arshad U. Hate speech detection on Twitter using transfer learning. Computer Speech & Language. 2022. https://www.sciencedirect.com/science/article/abs/pii/S0885230822000110.

6. Zalutska O., Molchanova M., Sobko O. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings. 2023. № 3387. P. 561–571.

7. Abusive Language Dataset. 2023. https://hatespeechdata.com/#Ukranian-header.

References

1. Mann S., Arora J., Bhatia M. Twitter Sentiment Analysis Using Enhanced BERT. Intelligent Systems and Applications. Springer. 2023. № 959. R. 263–271.

2. Slobodzian V., Molchanova M., Kovalchuk O. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. 12th International Conference on Advanced Computer Information Technologies. ACIT. 2022. R. 400–405.

3. European Court of Human Rights. Knowledge Sharing. 2023. https://www.echr.coe.int/Pages/home.aspx?p=caselaw/otherpublications&c=#n 1593094460135143431056 7_pointer.

4. Zabolotnia T. M., Sokolovska A. V. Metod avtomatyzovanoho vyznachennia naiavnosti obrazlyvoho vmistu tekstovykh povidomlen u sotsialnykh merezhakh. Visnyk ZhDTU. Seriia "Tekhnichni nauky". 2018. № 81. S. 103–108.

5. Ali R., Farooq U., Arshad U. Hate speech detection on Twitter using transfer learning. Computer Speech & Language. 2022. https://www.sciencedirect.com/science/article/abs/pii/S0885230822000110.

6. Zalutska O., Molchanova M., Sobko O. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings. 2023. № 3387. R. 561–571.

7. Abusive Language Dataset. 2023. https://hatespeechdata.com/#Ukranian-header.