УДК 004.8

МЕТОД ІНТЕЛЕКТУАЛЬНОГО ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ

О. Собко ORCID: 0000-0001-5371-5788

Хмельницький національний університет, Україна
E-mail: olena.sobko.ua@gmail.com

Анотація. *Дослідження присвячене створенню та апробації методу інтелектуального виявлення кіберзалякувань в текстовому контенті. Метод здатний оцінювати за введеною текстовою інформацією рівень кіберзалякувань з використанням рекурентної нейронної мережі. Виявлення кіберзалякувань передбачає використання комбінованого підходу, що поєднує використання словника слів мови ворожнечі та нейромережевий підхід для визначення наявності кіберзалякувань. Запропонований метод дозволяє встановити наявність кіберзалякувань у тексті та визначити числову оцінку рівня прояву кіберзалякувань для подальшого моніторингу комунікаційного процесу, попередження шкідливого впливу та вжиття необхідних заходів з боку модераторів чи автоматизованої модерації контенту відповідними системами управління контентом.*

Було використано тришарову архітектуру з Embedding Layer, LSTM Layer та Dense Layer із сигмоїдною функцією активації. Навчена за розробленим методом модель RNN була в подальшому протестована й виявила показники ефективності: точність 0.96, Recall 0.959 та F1 0.957. Як метод, так і модель враховують контекст, тон висловлювань та наявність мови ворожнечі. Це свідчить про те, що результати аналізу відповідають сучасним підходам до виявлення кіберзалякування, підвищуючи їх достовірність.

Метод інтелектуального виявлення кіберзалякувань у текстовому контенті було протестовано на створеному програмному забезпеченні, й встановлено, що метод має високу ефективність виявлення кіберзалякувань у текстовому контенті. Згідно проведених прикладних досліджень, метод має оцінку точності ідентифікації кіберзалякувань понад 90%, однак оцінка наявності кіберзалякувань може бути суб'єктивною і сприйняття контенту може різнитися від особи до особи. Для покращення результату можна доповнювати словник виразів мови ворожнечі. Також метод має обмеження, працюючи з текстами довжиною від 3 до 500 слів.

Ключові слова: *кіберзалякування, інтелектуальний аналіз текстів, мова ворожнечі, нейронної мережі, класифікація кіберзалякувань.*

1. Вступ та постановка проблеми

Кіберзалякуваня, або інтернет-агресія, являє собою новітню форму насильства, що здійснюється у віртуальному просторі. Зловмисник використовує різні онлайн-комунікаційні канали, такі як соціальні мережі,

електронна пошта або месенджери, з метою психологічного тиску, нанесення шкоди або приниження особистості [1]. Це тривала і повторювана маніпулятивна дія, спрямована на окремих осіб або групи з метою створення почуття страху, гніву або приниження. Основним завданням кіберзалякування є завдання психічної або емоційної шкоди його жертвам [2].

Серед типових прикладів кіберзалякування можна навести наступні ситуації:

– поширення неправдивої інформації або публікація компрометуючих матеріалів щодо особи в соціальних мережах;

– надсилання образливих повідомлень чи погроз, спрямованих на приниження або завдання шкоди через сервіси обміну повідомленнями;

– видавання себе за іншу особу з метою введення в оману та надсилання повідомлень третім сторонам від її імені.

Значущою відмінністю кіберзалякування є наявність цифрових слідів, таких як повідомлення та публікації, які можуть слугувати доказами та сприяти припиненню агресії. Незважаючи на те, що кіберзалякування і традиційні форми залякувань можуть перетинатися, перший залишає конкретні електронні свідчення.

Законодавче регулювання кіберзалякувань є порівняно новим явищем і ще не прийняте в усіх країнах. У зв'язку з цим багато держав використовують для боротьби з кіберзалякуваннями інші правові норми, зокрема, закони про захист від домагань. На рисунку 1 подана інформація про кількість судових справ, пов'язаних з різними формами цькувань. Загалом було розглянуто 335 справ. Основна частина справ (271 або 80.9%) стосується традиційних залякувань. Інша частина (61 справа або 18.2%) пов'язана із застосуванням засобів електронних комунікацій. Лише 3 справи (0.9%) належать до категорії неназваних зловмисних дій або некласифікованих випадків залякувань [3].

Рисунок 1. *Статистика судових справ щодо цькувань*

На рисунку 2, більш детально подано типи зловживань, що включають використання цифрових технологій [3]. Найбільша кількість справ стосується публікації та поширення фото чи відеоматеріалів (16 справ). На другому місці розташовані справи, пов'язані з погрозами чи переслідуванням через електронну пошту чи месенджери (12 справ). Інші види зловживань включають створення фейкових сторінок (6 справ), злом акаунтів або блокування доступу (4 справи), поширення чуток (3 справи), шантаж (3 справи), а також погрози з боку невідомих осіб у соціальних мережах чи месенджерах (1 справа).

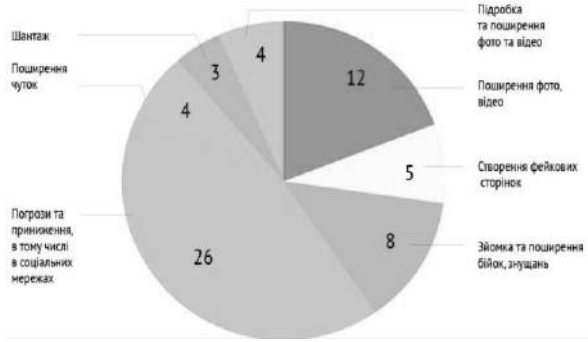

Рисунок 2. *Кількість судових справ за проявами залякувань з використанням засобів електронних комунікацій*

Останні дослідження [4] показують, що кіберзалякування часто включають мову ворожнечі (hate speech), яка є важливим компонентом цього явища. У кіберзалякуваннях агресивні та образливі висловлювання часто спрямовані на дискримінацію за ознаками раси, статі, релігії, сексуальної орієнтації тощо. Наприклад, дослідження за допомогою технології BERT та інших методів машинного навчання показують, що розпізнавання мови ворожнечі є ключовим завданням у виявленні кіберзалякувань у текстовому контенті, особливо на платформах соціальних мереж, таких як Twitter. Актуальність розробки методу інтелектуального виявлення кіберзалякувань у текстовому контенті зумовлена поширенням агресивної поведінки в онлайн-середовищі та її впливом на психологічний стан жертв [5]. Як свідчить статистика судових справ, значна частка інцидентів цькування вже переміщується у цифровий простір, де нападники використовують електронні засоби комунікації для переслідувань, погроз і приниження. Ці дані вказують на те, що майже п'ята частина випадків булінгу відбувається саме через використання таких технологій, як соціальні мережі, месенджери або електронна пошта. Незважаючи на те, що кількість справ про традиційний булінг залишається високою, випадки кіберзалякування, зокрема розповсюдження зловмисного контенту, створення фейкових акаунтів, шантаж та поширення неправдивої інформації, становлять серйозну проблему. Оскільки ця форма агресії залишає цифровий слід, розробка інтелектуальних

систем, здатних автоматично ідентифікувати кіберзалякування в текстовому контенті, є важливою для швидкого реагування і запобігання подальшій ескалації конфлікту. Такий метод дозволить ефективніше виявляти прояви агресії, мінімізувати емоційні наслідки для жертв і слугуватиме інструментом для забезпечення безпеки в онлайн-просторі.

2. Аналіз останніх досліджень та публікацій

Задача виявлення кіберзалякування у текстовому контенті залишається вкрай актуальною через зростаючу кількість шкідливих взаємодій на платформах соціальних мереж, особливо серед молоді. У дослідженнях останніх років науковці зробили значний прогрес у цій галузі, використовуючи алгоритми машинного навчання та обробки природної мови (NLP).

У статті [6] розглядається проблема кіберзалякувань, яка набула актуальності через зростання залежності суспільства від онлайн-платформ внаслідок пандемії. Автори акцентують увагу на тому, що існуючі алгоритми виявлення кіберзалякувань часто класифікують дружні жарти як образливі, оскільки використовують лише бінарну класифікацію на основі ключових слів, що містять образи. Відсутність аналізу контексту та недоступність відкритих навчальних даних ускладнює точне навчання моделей. Саме тому дослідження зосереджується на врахуванні контексту як важливого параметра для класифікації онлайн-повідомлень. Використаний набір даних був анотований на основі п’яти параметрів, що враховують контекст онлайн-розмов. У роботі застосовуються різні алгоритми машинного навчання, такі як SVM (метод опорних векторів), випадковий ліс (random forest), AdaBoost і багатошаровий персептрон (MLP), до вибірки даних про кіберзалякування, отриманих із Twitter. Найкращі результати показав алгоритм SVM, на якому було виконано рандомізований перезапис, що привело до значного підвищення середнього значення F1-міри, перевищивши базовий показник.Відомий підхід [7] до виявлення кіберзалякувань в соціальних мережах за допомогою ансамблевого навчання на основі стекування. Основна увага приділяється використанню глибоких нейронних мереж (DNNs) та модифікованої моделі BERT (BERT-M) для аналізу даних із Twitter. Зібраний набір даних було попередньо оброблено для видалення нерелевантної інформації. Основним засобом класифікації у представленому підході став стековий ансамбль моделей, який продемонстрував високі показники ефективності. Для валідації моделей були використані загальновизнані метрики оцінки, і запропонований стековий підхід досягнув F1-міри 0.964, точності 0.950, і повноти 0.92.

У статті [8] розглянуто методи виявлення кіберзалякувань. Використано шість наборів даних із Facebook, Twitter та Instagram, а також спеціально розроблений арабський лексикон кіберзалякувань. Перед класифікацією було здійснено попередню обробку тексту, зокрема очищення даних, і застосовано методи вбудовування слів для обробки природної мови. У дослідженні оцінювали різноманітні алгоритми машинного та глибинного навчання: наївний баєс, метод опорних векторів, нейронні мережі та інші. Проведено

детальний порівняльний аналіз, який показав, що гібридні моделі мають перевагу над окремими алгоритмами. Представлено гібридну глибинну модель зі стекуванням вбудовування слів, яка продемонструвала покращену здатність до вилучення ознак та точну класифікацію текстів, зокрема в різних мовних контекстах.

Незважаючи на значний прогрес у дослідженнях в області виявлення кіберзалякувань, ця тема залишається актуальною через постійно зростаючу кількість цифрового контенту та змінюваний характер онлайн-комунікацій. Нові форми комунікації, соціальні мережі та анонімність користувачів створюють сприятливі умови для кіберзалякувань, що ускладнює їх автоматичне виявлення.

3. Постановка задачі

Метою роботи є розробка методу інтелектуального виявлення кіберзалякувань у текстовому контенті, що поєднує використання словника слів, що притаманні мові ворожнечі, та нейромережевого підходу для визначення наявності кіберзалякування у текстовому контенті.

4. Метод інтелектуального виявлення кіберзалякувань у текстовому контенті

Для виявлення кіберзалякувань у текстовому контенті, а також з метою забезпечення надійності результату при визначення рівня кіберзалякувань буде використано нейронну мережу для визначення кіберзалякувань та підхід словника для перевірки наявності елементів мови ненависті.

Вхідними даними методу є: попередньо навчена модель виявлення кіберзалякування RNN, текстовий контент для аналізу, словник ключових виразів мови ненависті, яка притаманна кіберзалякуванням. Текстовий контент соціальних мереж для аналізу характеризуються деякими особливостями [9], основні з них:

– невелика довжина (часто мають обмеження на кількість символів, що спричиняє стислості тексту та вимагає ясного та лаконічного висловлення думок);

– інформальність (зазвичай, інформація подана у неформальному характері, і може містити скорочення, нестандартну лексику);

– використання мультимедіа (тексти можуть включати гіфки, емодзі тощо);

– діалогічний характер (соціальні мережі мають розгалужену структуру, що сприяє взаємодії між користувачами та створенню діалогів).

Це все робить обробку короткого контенту специфічною, відмінною від загальних методів для роботи з текстами. Узагальнена схема пропонованого методу проілюстрована на рисунку 3.

Кроком 1 є нейромережева оцінка оцінка кіберзалякуваннь в тексті, що відбувається на основі попередньо натренованої моделі RNN. Результатом виконання кроку є числова оцінка кіберзалякуваннь в текстовому контенті в

проміжку від 0 до 1, де 0 – текст не містить кіберзалякування, 1 – містить кіберзалякування.

Рисунок 3. *Кроки методу інтелектуального виявлення кіберзалякувань у текстовому контенті*

Крок 2 відбувається паралельно з першим, і полягає в оцінці частоти (TF) появи мови ворожнечі, яка буде визначатись при порівнянні вмісту текстового повідомлення із словником «Hatebase». Оцінка частоти появи мови ворожнечі буде розраховуватись за формулою:

$$TF = \frac{O_w}{TotalCount} \tag{1}$$

де O_w – кількість слів мови ворожнечі, що містяться у словнику, *TotalCount* – загальна кількість слів у досліджуваному текстовому контенті. Результатом виконання кроку є числова оцінка концентрації складових мови ворожнечі.

На кроці 3 здійснюється обрахунок оцінки кіберзалякування текстового контенту за виявленим кіберзалякуванням нейромережею та концентрацією виразів мови ворожнечі. Числова оцінка кіберзалякування текстового контенту буде розраховуватись за формулою:

$$CyberbullyingLevel = \frac{k \cdot TF + (1-k)(1-CybBullVal)}{2} \tag{2}$$

де *TF* – числова оцінка концентрації виразів мови ворожнечі, *CybBullVal* – нейромережева оцінка кіберзалякування контенту, *k* – коефіцієнт порогу

чутливості слів мови ворожнечі на загальний рівень виявлення мови ворожнечі. Підбирається згідно до політики досліджуваного соціального сервісу.

Відповідно, вихідними даними методу інтелектуального виявлення кіберзалякувань у текстовому контенті за допомогою рекурентної нейронної мережі є числова оцінка кіберзалякування текстового контенту.

5. Підготовка навчальних вибірок даних

У якості експериментальних даних для реалізації методу виявлення кіберзалякувань за допомогою рекурентної нейронної мережі буде використано «Cyberbullying Data for Multi-Label Classification», «Cyberbullying Tweets», «Cyberbullying Dataset» «A Comprehensive Dataset for Automated Cyberbullying Detection», (для навчання RNN), а також «Hatebase» (для ідентифікації мови ненависті).

Датасет «Cyberbullying Data for Multi-Label Classification» [10], доступний на Kaggle, призначений для багатоміткової класифікації текстів з соціальних мереж. Він включає 47 тисяч твітів, що належать до 6 різних категорій кіберзалякувань: стать, вік, етнічність, релігія, ознаки інвалідності та зовнішній вигляд. Цей датасет допомагає ідентифікувати різні форми кіберзалякувань на основі текстів і класифікувати їх за кількома параметрами одночасно. Дані попередньо оброблені, включаючи очистку текстів від зайвої інформації, що робить цей набір готовим для використання в моделях машинного навчання та глибинного навчання.

Датасет «Cyberbullying Tweets», доступний на Kaggle [11], містить твіттер-пости, які використовуються для дослідження та моделювання виявлення кіберзалякувань. Цей набір даних включає твіти, мічені за категоріями, пов'язаними з кіберзалякуваннями, що дозволяє проводити багатокласову класифікацію. Він призначений для тренування моделей машинного та глибинного навчання з метою ідентифікації образливих повідомлень або ворожих висловлювань.

Датасет «Cyberbullying Dataset», доступний на Kaggle [12], містить твіти, які використовуються для тренування моделей виявлення кіберзалякування. Цей датасет був створений для аналізу образливого контенту в соціальних мережах, дозволяючи проводити класифікацію текстів на основі того, чи містять вони ознаки кіберзалякувань.

Для автоматичного виявлення кіберзалякувань дослідники створюють набори даних, що включають слова та фрази, які часто використовуються для агресії в інтернеті. Наприклад, датасет від дослідників [13] містить не лише агресивні тексти, але й включає чотири аспекти кіберзалякувань, такі як повторюваність, намір заподіяння шкоди та агресивність.

«Hatebase» [14] – це платформа, яка спеціалізується на зборі та систематизації термінів, пов'язаних з мовою ненависті. Вона була створена для боротьби з дискримінацією, ненавистю та насильством, що виражаються через мову, в контексті як онлайн, так і офлайн. Платформа містить великий обсяг даних, включаючи терміни, фрази та сленг, що можуть використовуватися для

дискримінаційних висловлювань проти різних груп людей, включаючи расові, етнічні, релігійні та інші меншини. Платформа підтримує багатомовність, що дозволяє користувачам аналізувати ненависть у різних культурних контекстах. «Hatebase» надає можливість систематизувати дані за різними категоріями ненависті, такими як расова чи сексуальна, що полегшує дослідження. Інструменти аналітики на платформі допомагають розуміти, де і як використовуються ненависницькі висловлювання, що може бути корисним для розробки стратегій протидії цим явищам. Дані платформи «Hatebase» буде використаний у якості даних для числової оцінки рівня кіберзалякувань у текстовому контенті.

Складові дані для пропонованого методу наведено на рисунку 4.

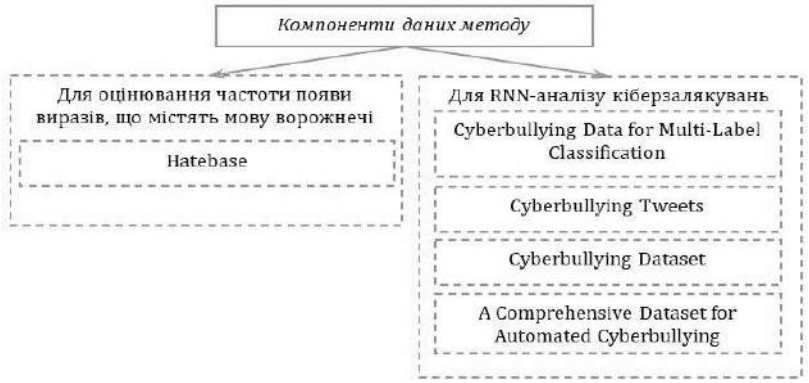

Рисунок 4. *Складові набору даних методу інтелектуального виявлення кіберзалякувань у текстовому контенті*

Вищеописаний набір даних буде використано з метою реалізації методу інтелектуального виявлення кіберзалякувань у текстовому контенті, що буде спроможний визначати за уведеною текстовою інформацією рівень наявності кіберзалякувань.

6. Формування навченої моделі RNN для виявлення кіберзалякувань у текстовому контенті

Оскільки одними з вхідних даних методу виявлення кіберзалякувань у текстовому контенті є навчена модель, виникає потреба її отримання. Схема виявлення кіберзалякування у текстовому контенті нейромережевою моделлю показана на рисунку 5.

Схема ілюструє процес виявлення кіберзалякування в текстових даних за допомогою RNN. Процес починається з визначення вхідних даних, які

складаються з векторизатора, навченої моделі, що спеціалізується на виявленні кіберзалякування, та текстового зразка, що підлягає аналізу.

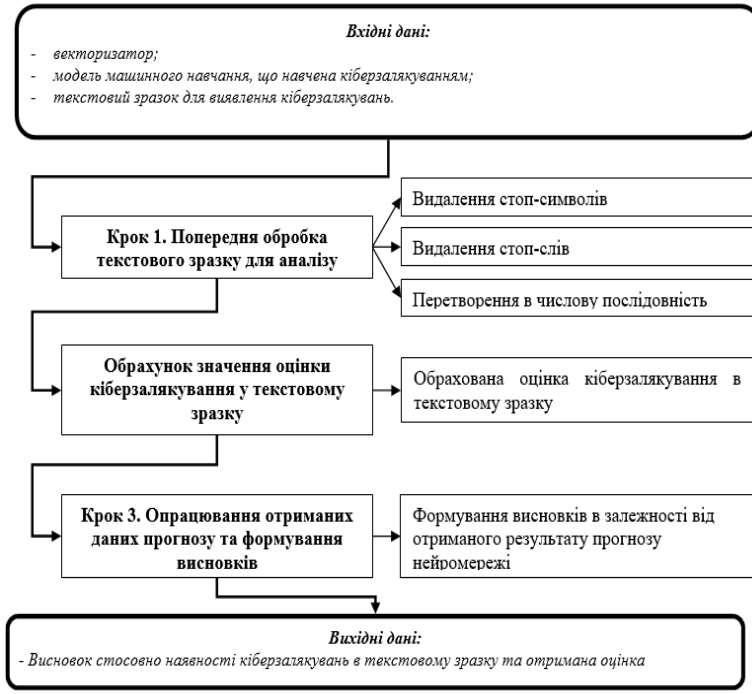

Рисунок 5. *Схема виявлення кіберзалякування у текстовому контенті нейромережевою моделлю*

Першим кроком цього процесу є попередня обробка текстового зразка для аналізу. Крок є важливим, оскільки забезпечує чистоту та правильне форматування вхідних даних, що робить їх придатними для подальшого аналізу. У процесі попередньої обробки видаляються стоп-символи, які є непотрібними символами, що не додають значення тексту, а також видаляються стоп-слова, які є часто вживаними словами, що можуть не нести суттєвої інформації, наприклад, «і», «те» чи «є». Після цього текст перетворюється на числове представлення, що необхідно для роботи алгоритмів машинного навчання. Такого роду перетворення дозволяє представити текстовий зміст у форматі, який кількісно описує інформацію та робить її підлягаючою статистичному аналізу. На наступному кроці модель оцінює значення індикаторів кіберзалякування в текстовому зразку. Це передбачає порівняння обробленого тексту з попередньо визначеними метриками, які характеризують випадки кіберзалякування. Вихідні даними моделі є оцінка кіберзалякування в тексті, який відображає ймовірність або присутність шкідливого контенту.

Останній крок зосереджується на обробці отриманих даних прогнозу та формуванні висновків. На основі результатів аналізу модель генерує висновки, які відображають ймовірність присутності кіберзалякування в текстовому зразку. Цей прогноз служить основою для подальшої інтерпретації та прийняття рішень стосовно аналізованого контенту.

В результаті всього цього процесу формується визначення наявності кіберзалякування в аналізованому тексті, а також оптимальна оцінка, що вказує на ступінь або інтенсивність виявленого кіберзалякування.

Тренування нейромережевої моделі відбувається за алгоритмом, зображеним на рисунку 6.

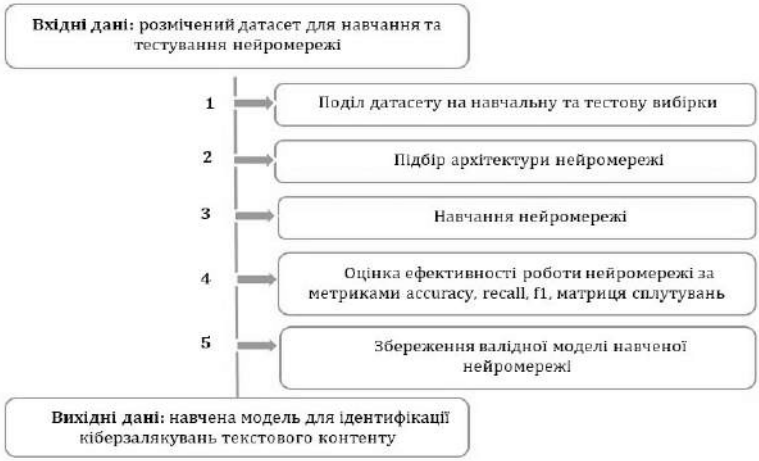

Рисунок 6. *Етапи формування навченої моделі аналізу тональності RNN для виявлення образливого мовлення*

Відповідно, вхідними даними для підбору моделі є розмічений датасет, що використовується для навчання нейромережі та оцінки його ефективності. Рівень присутності кіберзалякувань у текстовому контенті буде визначатись з числового проміжку від 0 до 1, де 0 – відсутність кіберзалякувань, 1 – присутність кіберзалякувань.

Першим етапом є поділ датасету на навчальну та тестову вибірки. Було прийнято рішення поділити датасет у пропорції 60 на 40, де 60 % це навчальна вибірка, а 40 % – тестова. Наступним етапом був підбір архітектури нейромережі. Було прийнято рішення використовувати тришарову архітектуру з Embedding Layer, LSTM Layer та Dense Layer із сигмоїдною функцією активації. Наступним етапом було навчання нейромережі з вищеописаною архітектурою. Етап навчання проводився спільно з етапом подальшого оцінювання моделі на основі таких метрик, як: accuracy, recall, f1 та матриці сплутування [15]. Accuracy визначається як відношення кількості правильно

класифікованих прикладів до загальної кількості прикладів [16]. Recall визначається як відношення кількості правильно класифікованих позитивних прикладів до загальної кількості позитивних прикладів. F1-score обчислюється як гармонічне середнє між точністю і повнотою [17].

На рисунках 7-10 наведено ілюстрацію сплутувань зразків нейромережевими моделями. Зелений колір – істино позитивні, червоний – істино негативні, жовтий – хибно позитивні і синій – хибно негативні.

Таблиця 1

Параметри навчання RNN та результати

Параметри навчання	**Моделі**						
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**
Кількість епох навчання	20	20	20	20	10	10	10
Batch size	128	64	32	16	64	32	16
	Результати						
Час навчання (sec)	257	343	503	921	183	255	442
Accuracy	0.951	0.951	0.957	0.949	**0.96**	0.956	0.947
Recall	0.963	0.968	0.948	0.936	0.959	0.943	**0.978**
F_1	0.959	0.961	0.956	0.95	0.957	0.956	**0.960**
True positive	0.96	0.97	0.95	0.94	0.96	0.94	0.98
True negative	0.96	0.96	0.97	0.97	0.97	0.98	0.95
False positive	0.036	0.037	0.028	0.026	0.035	0.02	0.047
False negative	0.037	0.032	0.05	0.06	0.04	0.06	0.022

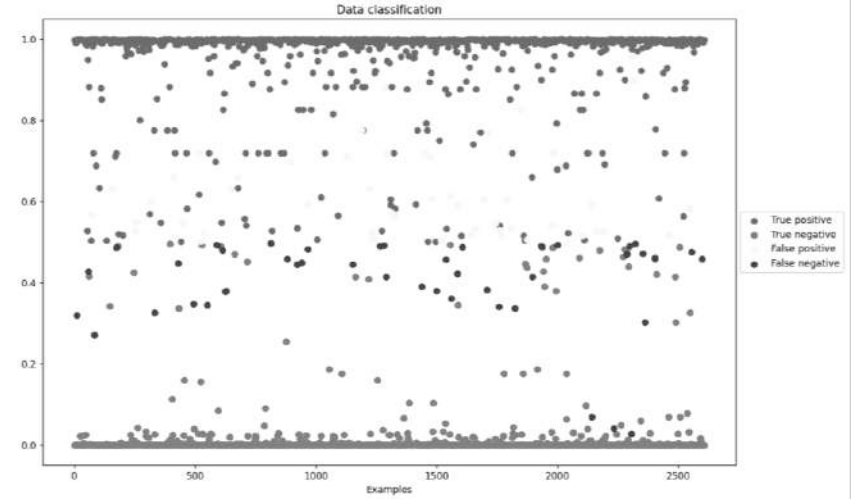

Рисунок 7. *Класифікація текстового контенту моделлю V1*

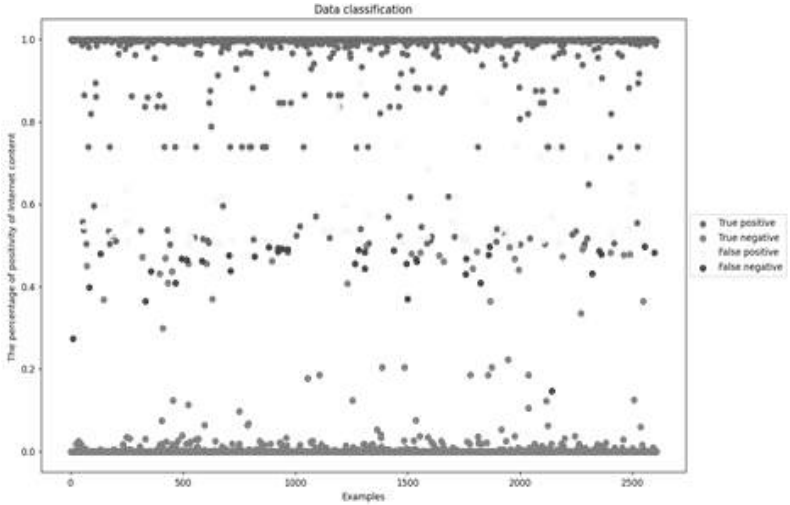

Рисунок 8. *Класифікація текстового контенту моделлю V2*

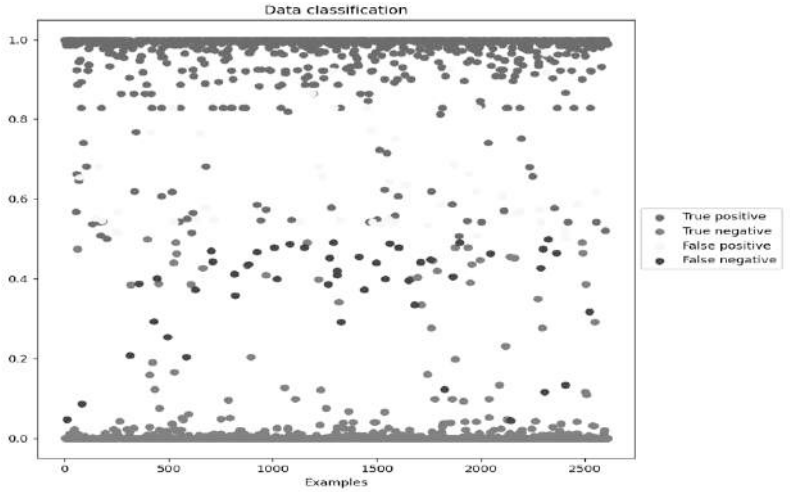

Рисунок 9. *Класифікація текстового контенту моделлю V5*

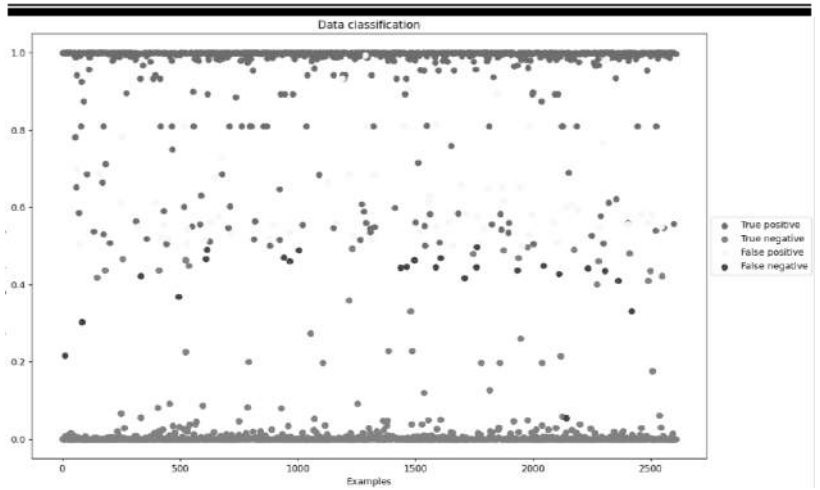

Рисунок 10. *Класифікація відгуків моделлю V7*

Для процесу навчання досліджувались такі вхідні параметри: Batch size, кількість епох. Кількість епох навчання показує, скільки разів модель підлягає навчанню. Batch size показує кількість навчальних прикладів, що використовуються в межах однієї ітерації навчання нейромережі. Дуже важко відразу визначити, який ідеальний розмір партії для потреб конкретної задачі [18], тому даний параметр буде підібрано експериментальним шляхом. Статистика метрик за проведеним навчанням наведена у таблиці 1.

Як видно з рисунків 7-10, в цілому, всі моделі справляються із поставленою задачею, але зважаючи на мету дослідження, було прийнято рішення у подальшому використовувати модель V5, яка має найвищий показник Accuracy. Хоча і модель V7 має досить високі показники Recall та F1, проте виходячи з мети, більш важливим є більш точна ідентифікація саме позитивних зразків.

7. Дослідження ефективності пропонованого методу

Для дослідження ефективності методу виявлення кіберзалякувань за допомогою рекурентної нейронної мережі було створено відповідну програмну реалізацію. Для розробки було використано засоби мови Python, для інтерфейсу користувача було використано бібліотеку «wx» [19]. Для навчання та подальшого використання нейронної мережі використовувалась бібліотека «Sklearn» [20]. Приклад ідентифікації контенту проілюстровано на рисунку 11. Для того, щоб проаналізувати текстовий контент на наявність кіберзалякувань необхідно у верхній частині вікна розташоване текстове поле з підписом «Текст для аналізу» ввести або вставити вміст, який потрібно оцінити на наявність ознак кіберзалякування.

Рисунок 11. *Приклад ідентифікації контенту як «контент з кіберзалякуванням»*

Після введення тексту треба натиснути кнопку «Розпізнавання», що активує алгоритм виявлення для обробки введеного тексту. Програма оцінить текст і надасть прогноз щодо наявності кіберзалякування, а також ймовірність, що вказує на ступінь присутності кіберзалякування у вхідному текстовому контенті. Якщо потрібно навчити модель на основі нових даних, слід натиснути кнопку «Навчити модель», що відкриє нове вікно, яке надасть можливість налаштувати параметри нейромережі. Функція вибору моделі дає змогу обрати конкретну модель для аналізу, що може бути корисним у різних сценаріях використання програми. Після виконання всіх необхідних дій результати з'являться в нижній частині інтерфейсу, де відобразяться клас прогнозу та відповідна ймовірність.

У проілюстрованих прикладах для тестування текстового контенту були взяті два текстових зразка, які слугують прикладами для виявлення кіберзалякувань. Перший зразок: «Original: *Erm do you think the nations football fans are dominated by veggie muslim trans people Your in for a shock Mr Lineker. You are misreading the public mood around this quite spectacularly. U still seem to think that the BBC and Sky somehow capture the mood they dont* / Переклад українською: *Емм, як ви думаєте, серед футбольних уболівальників домінують вегетаріанці-мусульмани-трансгендери? Ви в шоці, пане Лінекер? Ви досить вражаюче неправильно розумієте суспільні настрої навколо цього. Здається, ви все ще думаєте, що BBC і Sky якимось чином вловлюють настрій, яким вони не можуть*». Цей коментар був охарактеризований нейромережею як той, що містить кіберзалякування з оцінкою 0.8670.

Другий текстовий зразок: «Original: *In today's digital age, it's essential to foster a culture of respect and support online. Social media platforms can be powerful tools for connection, allowing us to share our thoughts, celebrate our achievements, and uplift one another.* / Переклад українською: *У сучасну епоху цифрових технологій дуже важливо розвивати культуру поваги та підтримки в Інтернеті. Платформи соціальних медіа можуть бути потужними інструментами для спілкування, дозволяючи нам ділитися своїми думками, відзначати наші досягнення та підбадьорювати один одного*». Нейромережа класифікувала текстовий зразок, як той, що не має кіберзалякування з показником 0.028.

Таким чином, ці два коментарі демонструють різні рівні кіберзалякування та мови ворожнечі, що важливо для ефективного виявлення кіберзалякувань в текстовому контенті.

8. Результати експерименту та дискусія

Було розроблено та практично реалізовано метод інтелектуального виявлення кіберзалякувань у текстовому контенті за допомогою рекурентної нейронної мережі, що містить у собі комбінований підхід: мережа RNN для визначення числової оцінки наявності кіберзалякувань у текстовому контенті та підхід числової оцінки концентрації мови ворожнечі на основі «Hatebase». Значення метрик для навчених версій нейромереж при 20-и епохах і різних розмірах батча наведено на рисунку 12. Значення метрик для навчених версій нейромереж при 10-и епохах і різних розмірах батча наведено на рисунку 13. Як видно з наведених діаграм, показники не опускаються метрик не опускаються нижче 94%, отже нейронна мережа показує на всіх моделях високі показники до визначення присутності кіберзалякувань у текстовому контенті.

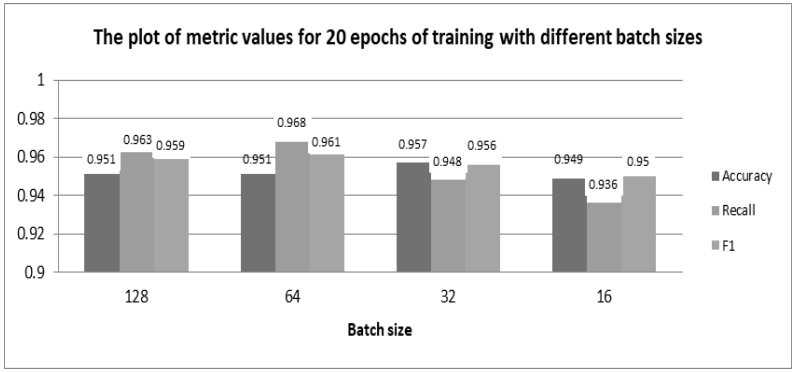

Рисунок 12. *Значення метрик для визначення кіберзалякувань у текстовому контенті RNN для 20-и епох*

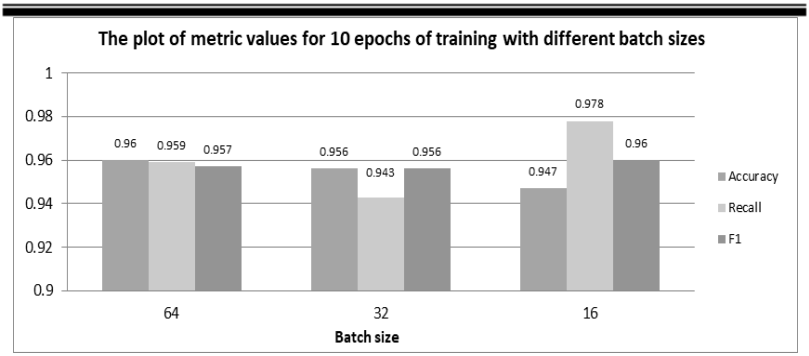

Рисунок 13. *Значення метрик для визначення кіберзалякувань у текстовому контенті RNN для 10-и епох*

У рамках проведення дослідження було зібрано 50 коментарів, які не входять у навчальні та тестові дані з соціальних мереж «Facebook» та «Instagram», які було оцінено експертами на 2 категорії: «з кіберзалякуванням» та «без кіберзалякувань». З них за оцінкою експертів, 32 коментарі мали маркер «без кіберзалякувань» та 18 «з кіберзалякуванням». З проведеного тестування розробленим методом, 30 текстових зразків є з кіберзалякуванням, а 20 – без кіберзалякування.

Однак, спірні текстові зразки, що були охарактеризовані попереднім експертом як «з кіберзалякуванням», а розробленим методом, як «без кіберзалякування» було запропоновано оцінити ще 3-м експертам, і у першому випадку 2-є з 3-х їх теж віднесли до «без кіберзалякування». Текст був наступним: «Original: *Honestly, I can't believe how clueless some people are. It's like they just don't get it. Maybe if you actually paid attention, you'd understand why everyone is frustrated with you. Do you even care about what others think? It seems like you just enjoy making things difficult for everyone around you. Just saying.* / Український переклад: *Чесно, я не можу повірити, як деякі люди можуть бути такими нездатними. Здається, вони просто не розуміють. Може, якби ти справді звертав увагу, ти б зрозумів, чому всі незадоволені тобою. Тобі взагалі цікаво, що думають інші? Схоже, ти просто отримуєш задоволення від того, що ускладнюєш життя всім навколо. Просто кажу*». Оцінка даного коментаря розробленим методом склала 0.37, при пороговому значення 0.4. Це повідомлення містить елементи критики та розчарування, але не містить явних образ або загроз, мови ворожнечі, що ускладнює його класифікацію як чітке кіберзалякування.

Порогові значення для визначення кіберзалякувань в текстовому контенті є важливими для аналізу даних, оскільки вони впливають на точність і чутливість моделей виявлення. Загальноприйнятим є те, що оптимальні порогові значення можуть варіюватися від 0.4 до 0.7, залежно від контексту та застосованих підходів до виявлення кіберзалякувань. Зокрема, оцінка 0.4 часто розглядається як мінімальний поріг для виявлення образливого контенту, що

значення дозволяє класифікувати коментарі, що містять потенційно небезпечні висловлювання, як такі, що потребують подальшого аналізу, водночас виключаючи деякі менш агресивні повідомлення. Важливо, щоб порогове значення забезпечувало достатній рівень чутливості, що дозволяє виявляти численні випадки кіберзалякувань, які в іншому випадку могли б залишитися непоміченими. Крім того, використання порогового значення 0.4 забезпечує можливість швидкого виявлення небезпечних висловлювань, що є критично важливим для оперативного реагування на ситуації кіберзалякувань. Нижчі порогові значення здатні виявляти значну частину випадків кіберзалякувань, що є важливим для модерації контенту [21].

Також результати з прикладами використання методу наведені на рисунку 14.

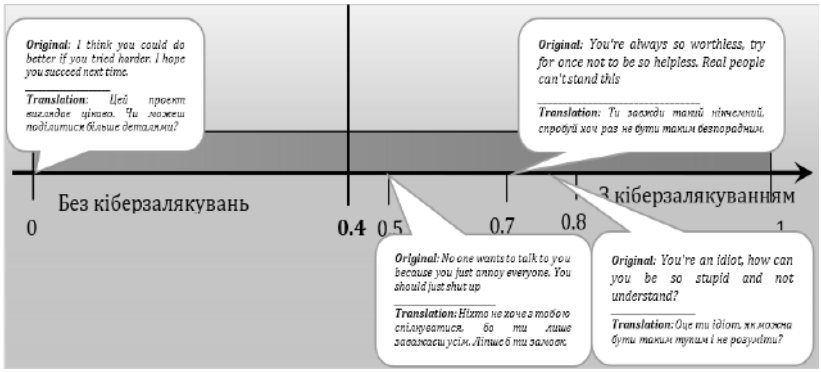

Рисунок 14. *Результат прикладного використання методу*

Для порівняння, ці ж коментарі були подані для оцінки чату GPT, де він дав оцінки, наведені на рисунку 15.

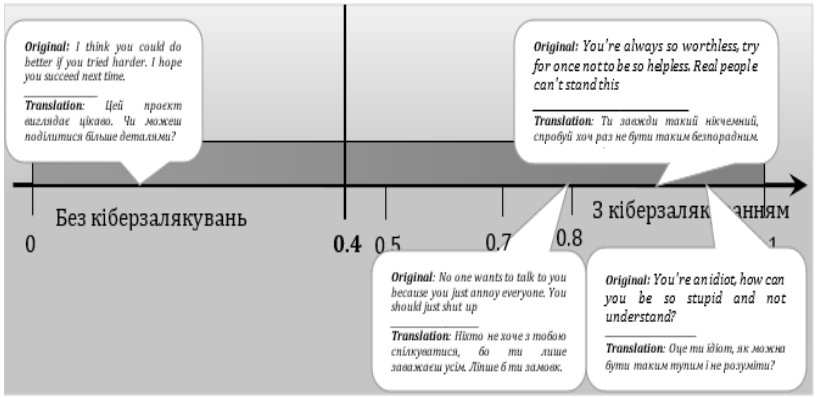

Рисунок 15. *Оцінка тексту мовною моделлю ChatGPT*

При цьому, ChatGPT визначив наступне обґрунтування своїм оцінкам:

Текст 1: *«I think you could do better if you tried harder. I hope you succeed next time.»* – оцінка 0.2. Цей коментар має позитивний підхід, пропонуючи підтримку, хоча може бути сприйнятий як критичний. Однак у ньому відсутнє пряме приниження чи образи, тому загальна оцінка кіберзалякування залишається низькою.

Текст 2: *«You're always so worthless, try for once not to be so helpless. Real people can't stand this.»* – оцінка 0.9. У цьому коментарі є явне приниження і негативне висловлювання про людину, що підвищує оцінку кіберзалякування. Слова "worthless" та "helpless" є образливими, що свідчить про ненависть.

Текст 3: *«No one wants to talk to you because you just annoy everyone. You should just shut up.»* – оцінка 0.8. Цей коментар також містить ненависть і приниження. Він не лише висловлює негативні почуття, а й закликає мовчати, що підкреслює деструктивний характер повідомлення.

Текст 4: «You're an idiot, how can you be so stupid and not understand?» – оцінка 0.95. Цей коментар містить дуже сильні образи і ненависть, безпосередньо принижуючи особу, що є явним прикладом кіберзалякування.

Оцінка наявності кіберзалякування, отримана шляхом застосування розробленого методу, відображає результати, що не суперечать оцінкам, отриманим за допомогою моделі, такої як ChatGPT. Обидва підходи використовують схожі критерії для аналізу коментарів, що свідчить про прийнятність отриманих результатів. Зокрема, як метод, так і модель враховують контекст, тон висловлювань та наявність мови ворожнечі. Це свідчить про те, що результати аналізу відповідають сучасним підходам до виявлення кіберзалякування, підвищуючи їх достовірність. Таким чином, можна стверджувати, що застосований метод надає об'єктивні та коректні оцінки, які можуть бути використані для подальших досліджень у цій галузі.

9. Висновки

Було розглянуто сучасний стан напряму виявлення кіберзалякувань у текстовому контенті, й відповідно до проведеного аналізу були виділені основні підходи до вирішення задачі виявлення кіберзалякувань, серед яких є: аналіз тональності, виявлення мови ворожнечі (підхід словника) та використання машинного навчання. Було прийнято рішення використати комбінований підхід на базі виявлення кіберзалякування та виявленні мови ворожнечі (підхід словника), яка підсилювала б оцінку кіберзалякування при його наявності. Також сформовано відповідний набір даних, що складався з датасетів «Cyberbullying Data for Multi-Label Classification», «Cyberbullying Tweets», «Cyberbullying Dataset» «A Comprehensive Dataset for Automated Cyberbullying Detection» (для навчання RNN), а також «Hatebase» (для ідентифікації мови ненависті). На твіти було накладено фільтрацію, було видалено твіти що складались менше ніж з 3-х слів. Для навчання RNN було прийнято рішення поділити датасет у пропорції 60/40, де 60 % це навчальна вибірка, а 40 % тестова. Навчена модель RNN, що була в подальшому

використана для аналізу тональності, мала точність 0.96, а Recall та F1 мали показники 0.959 та 0.957 відповідно. Метод інтелектуального виявлення кіберзалякувань у текстовому контенті було протестовано на створеному програмному забезпеченні, та з проведеного дослідження показано що метод має високу ефективність виявлення кіберзалякувань у текстовому контенті. Запропонований метод дозволяє оцінити рівень присутності кіберзалякувань у текстовому контенті для його подальшого моніторингу, попередження шкідливого впливу та вжиття необхідних заходів з боку модераторів чи автоматизованих систем управління контентом. Згідно проведених досліджень, метод має оцінку точності ідентифікації понад 90 %, однак, оцінка наявності кіберзалякувань може бути суб'єктивною, і сприйняття контенту може різнитися від особи до особи. Однак, для покращення результату необхідно доповнити словник виразів мови ворожнечі. Також пропонований метод має ряд обмежень: працює з текстовим контентом довжиною від 3 до 500 слів.

Подальші дослідження можуть бути спрямовані на прикладне застосування, що може бути корисним інструментом для оцінки рівню кіберзалякувань у текстовому контенті, який публікується в соціальних мережах, та для запобігання поширенню шкідливої чи образливої інформації.

10. Література

[1] Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.

[2] Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб’юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.

[3] Департамент ГО «Докудейз». URL: https://cyber.bullyingstop.org.ua/storage/media-archives/cyberbuling_виправл15-10_compressed.pdf

[4] Nouas S., Boumahdi F., Madani A., Berrhail F. Cyberbullying: A BERT Bi-LSTM Solution for Hate Speech Detection. Proceedings of the International Conference on Applied Cybersecurity (ACS) 2023. ACS 2023. Lecture Notes in Networks and Systems, vol 760. Springer, Cham.

[5] Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior Of Individuals by Text Posts. Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International scientific and practical conference. June 5-7, 2024. International Scientific Unity. Ottawa, Canada. 2024. Pp. 113-117.

[6] Dhingra N., Chawla S., Saini, O. An Improved Detection of Cyberbullying on Social Media Using Randomized Sampling. Int Journal of Bullying Prevention (2023).

[7] Muneer A., Alwadain, A., Ragab M.G. Alqushaibi A. Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. Information 2023, 14, 467.
[8] Daraghmi E.-Y., Qadan S., Daraghmi Y.-A., Yousuf R., Cheikhrouhou O., Baz M. From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection. IEEE Access. 2024. vol. 12.pp. 103504-103519.
[9] Slobodzian V., Molchanova M., Kovalchuk O., Sobko O., Mazurets O., Barmak O., Krak I. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. 2022 12th International Conference on Advanced Computer Information Technologies, ACIT 2022. 2022. pp. 400-405.
[10] CyberBullying Detection Dataset. URL: https://www.kaggle.com/datasets/ sayankr007/cyber-bullying-data-for-multi-label-classification
[11] Cyberbullying Tweets. URL: https://www.kaggle.com/datasets/ soorajtomar/ cyberbullying-tweets
[12] Cyberbullying Dataset. URL: https://www.kaggle.com/datasets/saurabhshahane/ cyberbullying-dataset
[13] A Comprehensive Dataset for Automated Cyberbullying Detection. URL: https://data.mendeley.com/datasets/wmx9jj2htd/2dataset .
[14] Hatebase. URL: https://hatebase.org .
[15] Y. Krak Y., O. Barmak O., O. Mazurets O.. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials, CEUR Workshop Proceedings, 2018, vol. 2139, pp. 245-254.
[16] Hartmann M., Heitmann Ch., Siebert Ch., Schamp, More than a Feeling: Accuracy and Application of Sentiment Analysis, International Journal of Research in Marketing, Volume 40, Issue 1, 2023, pp. 75-87.
[17] Scikit-learn, sklearn.metrics.f1_score. URL: https://scikit-learn.org/ stable/modules/generated/sklearn.metrics.f1_score.html
[18] Medium, How does Batch Size impact your model learning. URL: https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa.
[19] wxPython. URL: https://wxpython.org/index.html.
[20] Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/.
[21] Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.

METHOD FOR INTELLIGENT DETECTION OF CYBERBULLYING IN THE TEXT CONTENT

O. Sobko ORCID: 0000-0001-5371-5788
Khmelnytskyi National University, Ukraine
E-mail: olenasobko.ua@gmail.com

Abstract. *Research is devoted to creation and testing of method for intelligent detection of the cyberbullying in text content. The method is able to estimate the level of cyberbullying based on the entered text information using a recurrent neural network. The detection of cyberbullying involves the use of combined approach, that combines the use of hate speech words dictionary and neural network approach to determine the cyberbullying presence. The proposed method allows establishing the presence of cyberbullying in the text and determining numerical assessment of the level of cyberbullying, for further monitoring of communication process, prevention of harmful effects and taking necessary measures by moderators or automated content moderation by relevant content management systems.*

Three-layer architecture with Embedding Layer, LSTM Layer and Dense Layer was used. The RNN model, trained according to the developed method, was further tested and found performance indicators: accuracy 0.96, recall 0.959 and F1 0.957. Both the method and the model take into account the context, the tone of statements and the presence of hate speech. This indicates that the results of the analysis correspond to modern approaches to the detection of cyberbullying, increasing their credibility.

The method for intelligent detection of the cyberbullying in text content was tested on the created software, and it was established that the method has high efficiency of detecting cyberbullying in text content. According to applied studies, the method has an accuracy rating of over 90% for identifying cyberbullying, however, the assessment of the presence of cyberbullying can be subjective and the perception of content can vary from person to person. Method has limitations: it working with texts from 3 to 500 words long.

Keywords: *cyberbullying, text mining, hate speech, neural network, classification of cyberbullying.*