DOI 10.31891/2307-5732-2024-333-2-32
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
КЛІМЕНКО ВАЛЕРІЯ
Хмельницький національний університет
https://orcid.org/0000-0001-5869-4269
e-mail: ler.klimenko.8@gmail.com
АНДРОЩУК ВЛАДИСЛАВ
Хмельницький національний університет
e-mail: vladandroschuk0@gmail.com

МЕТОД НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ КІБЕРБУЛІНГУ З ВИКОРИСТАННЯМ ХМАРНИХ СЕРВІСІВ ТА ОБ'ЄКТНО-ОРІЄНТОВАНОЇ МОДЕЛІ

У роботі пропонується практичний підхід до виявлення кібербулінгу із використанням нейронної мережі BiLSTM, навченої за допомогою хмарних сервісів та застосунку, що реалізовує запропонований метод виявлення кібербулінгу. Для дослідження ефективності запропонованого методу було створено об'єктно-орієнтовану програмну реалізацію середовища програмування PyCharm, а також ноутбук для виконання в хмарному сервісі «Google Colab» для навчання нейромережі.

Ключові слова: виявлення кібербулінгу, нейронна мережа, об'єктно-орієнтована модель, кібербулінг, хмарні сервіси, Google Colab, BiLSTM, твіт

MOLCHANOVA MARYNA, MAZURETS OLEKSANDR, SOBKO OLENA, KLIMENKO VALERIIA, ANDROSCHUK VLADISLAV
Khmelnytskyi National University

METHOD FOR CYBERBULLYING NEURONETWORK DETECTION USING CLOUD SERVICES AND OBJECT-ORIENTED MODEL

The article proposes the method for detecting cyberbullying in posts of social Internet networks, designed for the automated detection of cyberbullying in text messages published in social networks using a neural network approach. The current state of the field of automated detection of cyberbullying is considered, where, based on the studied material, it is proposed to create a method of neural network detection of cyberbullying using cloud services and an object-oriented model. To investigate the effectiveness of the proposed method, a software implementation written in the Python programming language in the PyCharm programming environment was created, as well as a laptop for execution in the Google Colab cloud service for neural network training. The English-language "Cyberbullying Classification" dataset consisting of 39,747 samples was used as research data, from which 8,000 samples containing cyberbullying and 7,945 samples without cyberbullying were selected.

The proposed approach based on the object-oriented model contributes to the creation of flexible, adaptive and easily expandable systems, and the use of cloud services for training neural networks provides an opportunity not to be limited exclusively to the resources of a physical personal computer, but to use specialized computing resources, such as graphics processors (GPU) or tensor processors (TPU), which significantly accelerate the process of training neural networks. The approach has some limitations, the main limitation of the proposed approach is the maximum length of the input sequence, which is 500 characters. Considering the average length of a tweet is about 200 characters, such a limitation is reasonable. In order to improve the result, you can supplement the training dataset and specify some restrictions on the minimum and maximum length of records.

Keywords: cyberbullying detection, neural network, object-oriented model, cyberbullying, cloud services, Google Colab, BiLSTM, tweet

Аналіз предметної області

У контексті сучасних тенденцій, коли соціальні мережі стають необхідною складовою соціального життя, проблема кібербулінгу стає зростаючою та надзвичайно актуальною. Збільшення кількості користувачів у цифровому просторі створює сприятливе середовище для поширення ворожих або образливих повідомлень, які можуть негативно впливати на психічний стан та емоційний стан людини.

Нейромережеві моделі покликані розпізнавати ключові ознаки кібербулінгу, такі як агресивна лексика, образливі коментарі та загрози, забезпечуючи автоматизовану систему виявлення без необхідності ручного моніторингу. Здатність автоматично аналізувати значні обсяги інформації за короткий проміжок часу дозволяє вчасно виявляти та реагувати на можливі випадки кібербулінгу. Застосування нейромережевих технологій дозволяє підвищити ефективність та точність цього процесу, що важливо для забезпечення

безпеки та цифрового благополуччя користувачів соціальних мереж. Використання хмарних сервісів для навчання нейромережевих класифікаторів дозволяє працювати з різноманітними моделями та експериментувати з різними підходами до навчання нейромереж, не обмежуючись ресурсами локального комп'ютера, а об'єктно-орієнтована модель розробленого застосунку надає можливість подальшої розширюваності базового функціоналу.

Кібербулінг, також відомий як інтернет-мобінг, є формою агресії, що відбувається в онлайн-середовищі. Цей вид нападу характеризується використанням різноманітних засобів комунікації, таких як соціальні мережі, електронна пошта та месенджери, з метою тиснути, завдавати шкоди та принижувати цільову особу чи групу [1]. Кібербулінг є систематичним та повторюваним видом маніпуляційної поведінки, спрямованою на індивіда чи групу, з метою залякування, викликання гніву або приниження. Основна мета кібербулінгу полягає в спричиненні психологічних або емоційних травм жертві [2].

Кібербулінг може виявлятися у формі пропаганди війни, національної та релігійної ворожнечі, або сприяти зміні конституційного ладу чи територіальної цілісності України. Він також може поширювати фашизм та неофашизм, образити націю чи окремі особистості за національною ознакою, а також пропагувати бузувірство, блюзнірство та проявляти неповагу до національних і релігійних святинь. Крім того, кібербулінг може призводити до приниження особистості, включаючи знущання з фізичних вад чи каліцтва, а також з душевнохворих та літніх людей. Він також може пропагувати невігластво та виявляти неповагу до батьків. Характеристики кібербулінгу включають систематичні та повторювані дії, що можна визначити як акти кібербулінгу. Вони передбачають наявність різних сторін: кривдника (булі), потерпілого (жертви кібербулінгу) та спостерігачів, якщо такі існують. Дії або бездіяльність кривдника призводять до заподіяння психічної та/або фізичної шкоди, приниження, страху, тривоги, підпорядкування інтересам кривдника та/або призводять до соціальної ізоляції потерпілого. До найпоширеніших видів кібербулінгу належать такі види, як: використання особистої інформації, анонімні погрози, хепіслепінг, обмовляння або зведення наклепів, переслідування, тролінг, флеймінг, онлайн-грумінг, секстинг [3].

З проведеного аналізу, запропоновано автоматизувати виявлення кібербулінгу в дописах соціальних інтернет-мереж. Автоматизація виявлення кібербулінгу сприятиме створенню безпечних вебсередовищ.

Останні публікації

Проблемою виявлення кібербулінгу наразі займається велика кількість науковців по всьому світу. Європейський суд з прав людини (CEDH), дає визначення кібербулінгу, як «усі форми вираження, усні чи письмові, які поширюють, підбурюють, сприяють чи виправдовують ненависть на ґрунті нетерпимості» [4].

Метою дослідження [5] є ідентифікація коментарів із кібербулінгом. Експериментальні дані беруться з веб-сайту машинного навчання «Kaggle» та «Github». Вченими запропоновано використовуються три різних алгоритми машинного навчання: логістична регресія, метод опорних векторів, багатокласовий наївний Баєс для визначення найкращого алгоритму машинного навчання за метриками оцінки для передбачення коментарів. Основна мета дослідження полягає в передбаченні коментарів із кібербулінгом та їх сили, такої як легкий, сильний, помірний.

Інформаційно-комунікаційні технології збагатили соціальні мережі та полегшили комунікацію. Проте кібербулінг на цих платформах мав серйозні наслідки. Механізми, що залежать від користувача, такі як звітність, блокування та видалення повідомлень з кібербулінгом онлайн, є ручними і неефективними. Представлення тексту «bag-of-words» без метаданих обмежує класифікацію тексту з кібербулінгом в кіберпросторі. У дослідженні [6] авторами розроблено автоматизовану систему виявлення кібербулінгу двома підходами: традиційним машинним навчанням та передавальним навчанням. Авторами використані дані AMiCA, які включають значну кількість контексту кібербулінгу та структурований процес анотації. Текстові, сентимент та емоційні, статичні та контекстуальні вбудовування слів, психолінгвістичні, термінологічні та токсичні функції були використані в традиційному підході машинного навчання. Використання комбінації текстових, сентиментальних, вбудовувань DistilBert, психолінгвістичних та токсичних функцій у модель логістичної регресії дозволило підняти її ефективність до значення F-міри 64,8%, перевершивши Linear SVC за швидкістю тренування та ефективністю обробки високоінформаційних функцій. Використаний підхід передавального навчання полягав у доналаштуванні оптимізованих версій попередньо навчених мовних моделей, а саме DistilBert, DistilRoBerta та Electra-small, які виявилися більш швидкими у вирахуванні тренувань, ніж їх базові форми. Доналаштована нейромережа DistilBert мала найвище значення F-міри на рівні 72,42%, перевищивши традиційний метод машинного навчання. Авторами встановлено, що передавальне навчання було найкращим для підвищеної ефективності та зменшення зусиль, оскільки воно вимагало відсутності інженерії ознак та попередньої обробки даних.

Метою роботи є створення методу нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі.

Основна частина

Метод виявлення кібербулінгу в дописах соціальних інтернет-мереж призначений для використання нейромережевого підходу з метою автоматичного виявлення кібербулінгу в текстових повідомленнях, які публікуються у соціальних мережах. Схема та кроки методу наведені на рис. 1.

Вхідними даними для методу виявлення кібербулінгу в дописах соціальних мереж є векторизований текст, який був оброблений під час тренування нейромережі, а також збережена навчена модель для виявлення кібербулінгу та допис користувача для аналізу наявності кібербулінгу.

Перший крок включає передпопередню обробку допису для аналізу, що полягає в вилученні зайвих символів та слів, таких як розділові знаки та спеціальні символи, а також перетворення його у числову послідовність з використанням збереженого векторизатора. Ця обробка допомагає зменшити обсяг даних та підготувати їх для подальшого аналізу за допомогою збереженої нейромережевої моделі BiLSTM.

Другий крок включає застосування нейромережі для оцінки наявності кібербулінгу у дописі. Нейромережа отримує на вході текстові дані (перетворені на попередньому етапі у числову послідовність) та повертає прогноз щодо наявності кібербулінгу в тестовому дописі соціальної мережі. Архітектура нейромережі наведена на рисунку 2.

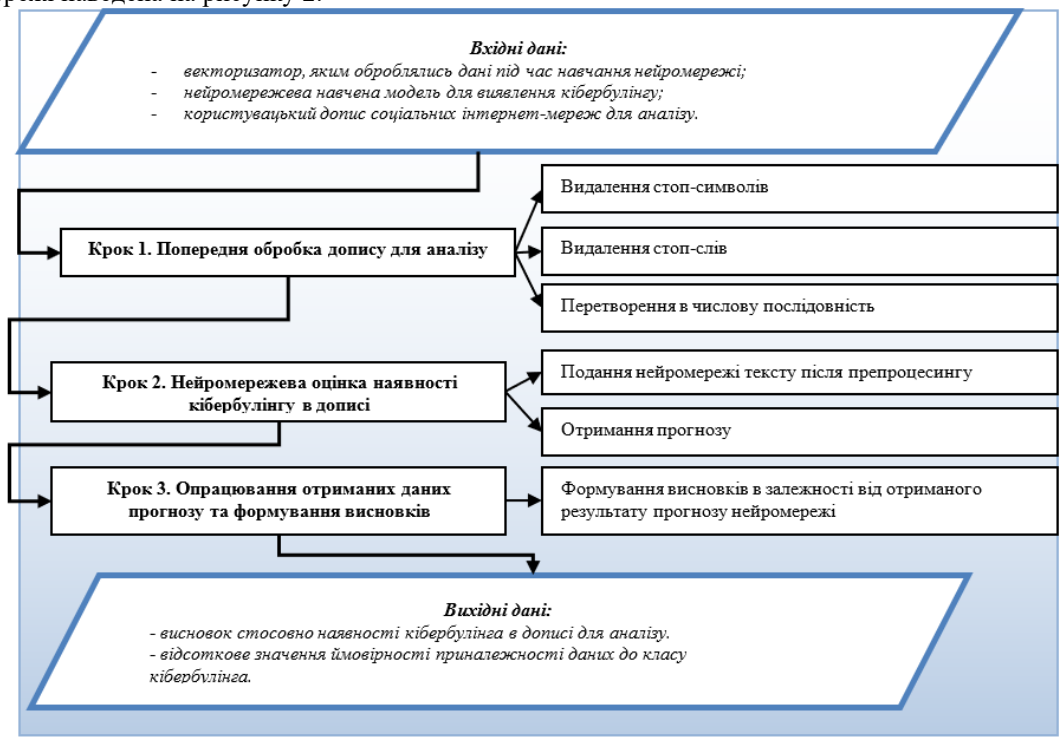


Рис. 1. Схема методу виявлення кібербулінгу в текстових повідомленнях

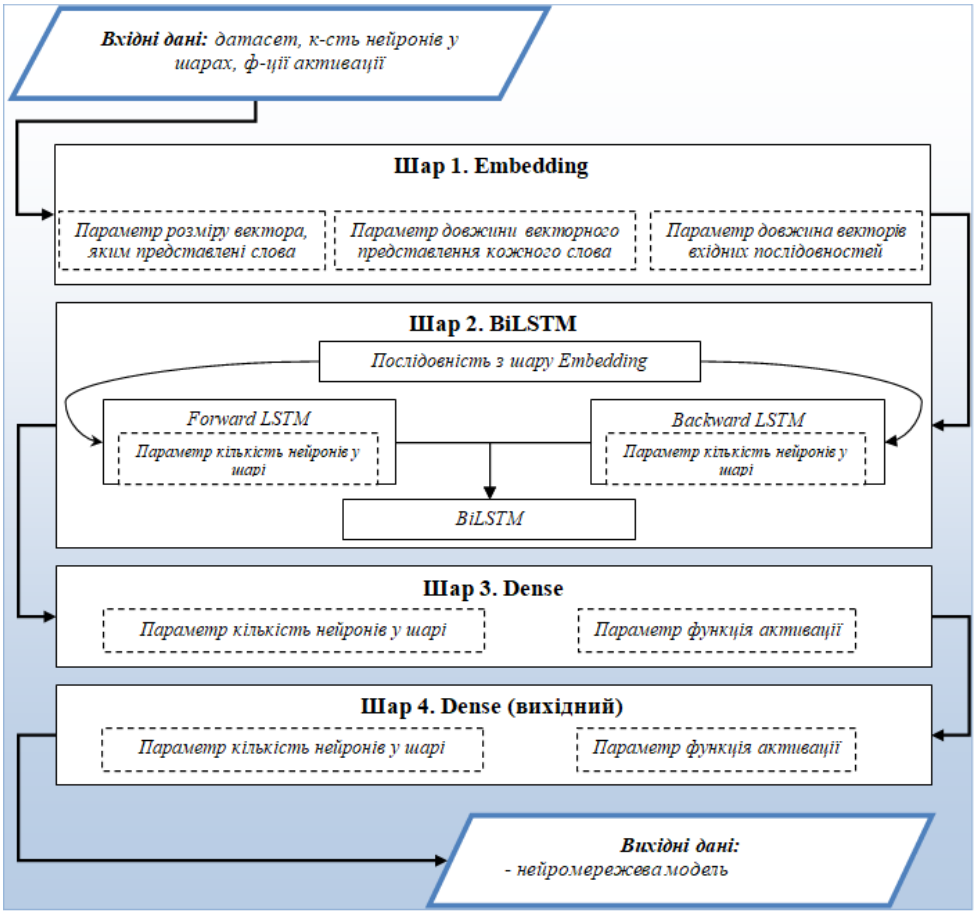


Рис. 2. Архітектура нейромережевої моделі BiLSTM

Архітектура нейронної мережі включає чотири основні шари: Embedding, який перетворює слова або символи у вектори фіксованої довжини; BiLSTM, що об'єднує виходи з обох напрямків для використання контекстної інформації з обох сторін вхідної послідовності; і два повнозв'язних шари Dense. Модель потребує навчання для забезпечення можливості виявлення кібербулінгу в дописах соціальних мереж.

Третім кроком є обробка отриманих прогнозних даних та формування висновків для користувачів. Висновок містить відповідь щодо наявності або відсутності кібербулінгу, а також відсоткову оцінку контенту, що містить кібербулінг, та того, що не містить його.

Результатом роботи методу є висновок про наявність кібербулінгу в аналізованому дописі та відсоткове значення ймовірності його належності до класу кібербулінгу.

Для оцінки ефективності методу розроблено об’єктно-орієнтовану інформаційну систему виявлення кібербулінгу, наведену на рис. 3, яка складається з усіх необхідних етапів і компонентів для реалізації та апробації запропонованого методу. Архітектура включає чотири підсистеми: «Підсистема навчання нейронної мережі», «Підсистема препроцесування», «Підсистема аналізу датасету» та головна «Підсистема виявлення кібербулінгу», а також набір даних і навчену нейронну мережу BiLSTM.

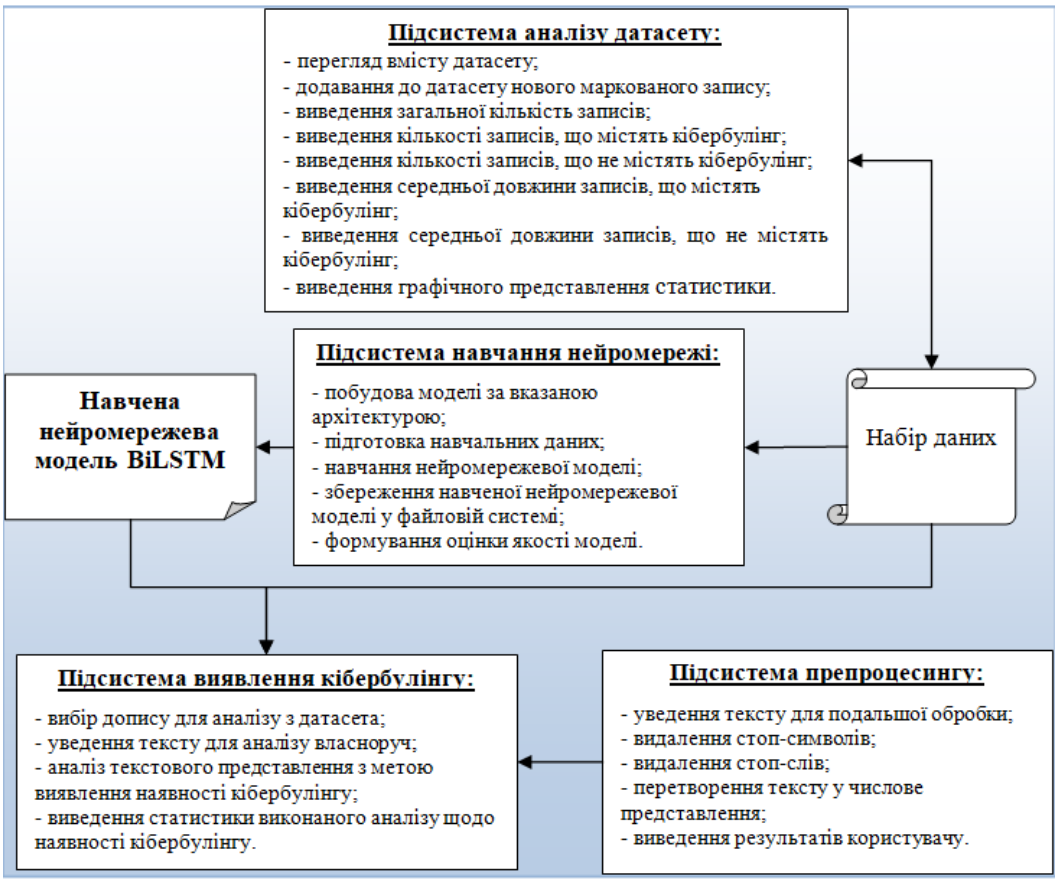


Рис. 3. Проектна архітектура інформаційної системи виявлення кібербулінгу

Таким чином, запропонований метод виявлення кібербулінгу в текстових повідомленнях дозволяє здійснювати виявлення кібербулінгу в дописах соціальних інтернет-мереж.

Дані дослідження

Для навчання нейромережі BiLSTM буде використано набір даних «Cyberbullying Classification» [6]. Сформований набір даних містить близько 40 000 твітів, позначених відповідно до класу кібербулінгу: вік, етнічна приналежність, стать, релігія, інші види кібербулінгу, не кібербулінг. Дані англомовні та збалансовані, містять близько 8000 твітів кожного класу. Проте у рамках дослідження буде використано тільки 2 класи – кібербулінг , який об’єднає всі види кібербулінгу та не кібербулінг [7]. Результуючий датасет налічує 39747 зразків, з яких буде обрано 8000 зразків, що містять кібербулінг, та 7945 записів без кібербулінгу. Набір даних є англомовним.

Дослідження ефективності методу виявлення кібербулінгу

Проведення експериментального дослідження для оцінки ефективності розробленого методу виявлення кібербулінгу в соціальних мережах є ключовим кроком. Для цього було створено програмну реалізацію, яка складається із ноутбука реалізованого у хмарному сервісі «Google Colab», що призначений для навчання моделей нейромережі BiLSTM та збереження їх для подальшого використання як вхідних моделей у десктопному застосунку для виявлення кібербулінгу. Десктопний застосунок реалізовано засобами мови Python у середовищі розробки PyCharm. Дослідження опирається на метрики точності (Accuracy) та F1-міри. Також проведено аналіз навчальних і валідаційних даних для оцінки роботи системи.

Були проведені експерименти з навчання нейромережі, які мали на меті сприяти покращенню базового відсотку ідентифікації, зміни стосувались кількості нейронів в шарах, кількості епох навчання та довжини словника і вхідних послідовностей [8]. Результати експерименту наведені в таблиці 1. По результатам експерименту побудовано графік (рис. 4), що ілюструє значення метрик для вказаних в таблиці 1 варіантів дослідження.

Таблиця 1

Параметри нейромережі та результати за метриками

Параметри нейромережі:	Accuracy	Loss	F1-міра
К-сть епох: 5. К-сть нейронів в шарах LSTM та Dense: 64. Розмір словника: 5000 (BiLSTM 1)	0.90	0.17	0.89
К-сть епох: 7. К-сть нейронів в шарах LSTM та Dense: 64. Розмір словника: 5000 (BiLSTM 2)	0.91	0.14	0.9
К-сть епох: 15. К-сть нейронів в шарах LSTM та Dense: 64. Розмір словника: 5000 (BiLSTM 3)	0.92	0.12	0.91
К-сть епох: 25. К-сть нейронів в шарах LSTM та Dense: 64. Розмір словника: 5000 (BiLSTM 4)	0.96	0.11	0.92
К-сть епох: 15. К-сть нейронів в шарах LSTM та Dense: 128. Розмір словника: 5000 (BiLSTM 5)	0.93	0.12	0.92
К-сть епох: 15. К-сть нейронів в шарах LSTM та Dense: 32. Розмір словника: 5000 (BiLSTM 6)	0.94	0.15	0.91
К-сть епох: 15. К-сть нейронів в шарах LSTM та Dense: 32. Розмір словника: 3000 (BiLSTM 7)	0.91	0.18	0.87
К-сть епох: 15. К-сть нейронів в шарах LSTM та Dense: 128. Розмір словника: 3000 (BiLSTM 8)	0.92	0.16	0.91

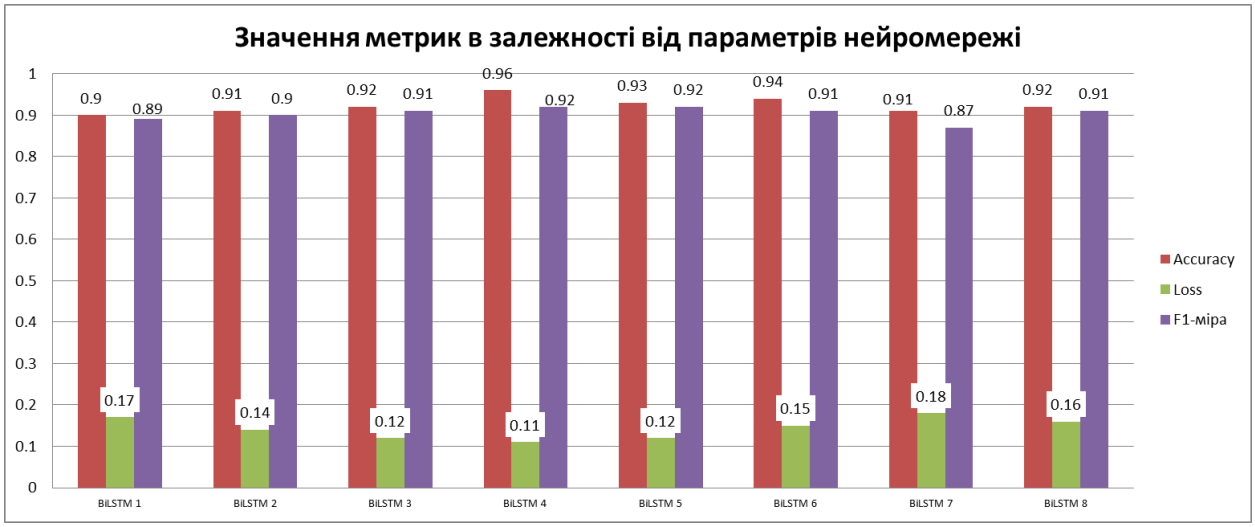


Рис. 4. Графік значень метрик в залежності від параметрів нейромережі

З таблиці 1 та графіку на рис. 4 видно, що всі параметри майже однакові, але версія BiLSTM з 25 епохами, 64 нейронами у шарах LSTM та Dense та розміром словника 5000 показала кращий результат, досягнувши метрики точності 0.96 та F1-міри 0.92. Функція втрат склала 0.11.

Однак результати на валідаційній вибірці були трохи гіршими: точність становила 0.86, а F1-міра – 0.83. Далі було проаналізовано датасет для визначення довжини твітів. Показано, що середня довжина твітів із кібербулінгом була 146.88 символів, тоді як середня довжина твітів без кібербулінгу становила 83.11 символів. Розподіл твітів за довжиною для категорії «Cyberbullying» та «Not cyberbullying» наведено на рис. 5.

Як видно з розподілів твітів за довжиною, твіти що містять кібербулінг, як правило, довші. У той час як твіти без кібербулінгу здебільшого не перевищують значення в 250 символів. Тому наступним кроком експерименту було відділення усіх твітів, що складають менше 6-и слів, а також відділення твітів, що складаються із довжини більше 1000 символів. Така фільтрація дещо зменшила розмірність вибірки, і вона стала складатись із 37995 твітів категорії «Cyberbullying» та 6884 твітів категорії «Not cyberbullying». Після використання даних маніпуляцій над вибіркою, нейромережу було знову перенавчено.

Як видно з рис. 6 (а) та 6 (б), результати навчання покращились, і на 15-ти епохах точність склала 0.97, у той час як в експерименті без фільтрації даних точність склала 0.92, при функція втрат 0.05, на ряду з значенням 0.12 до фільтрації. Перевіривши дану модель на валідаційних даних, метрики також покращили значення, склавши 0.91 за метрикою Accuracy та 0.89 за метрикою F1-міри.

Виконавши перенавчання на 25 епохах, показники дещо зросли і склали 0.989 за метрикою Accuracy та 0.95 за метрикою F1-міри на навчальному наборі та 0.928 за метрикою Accuracy та 0.912 за метрикою F1-міри.

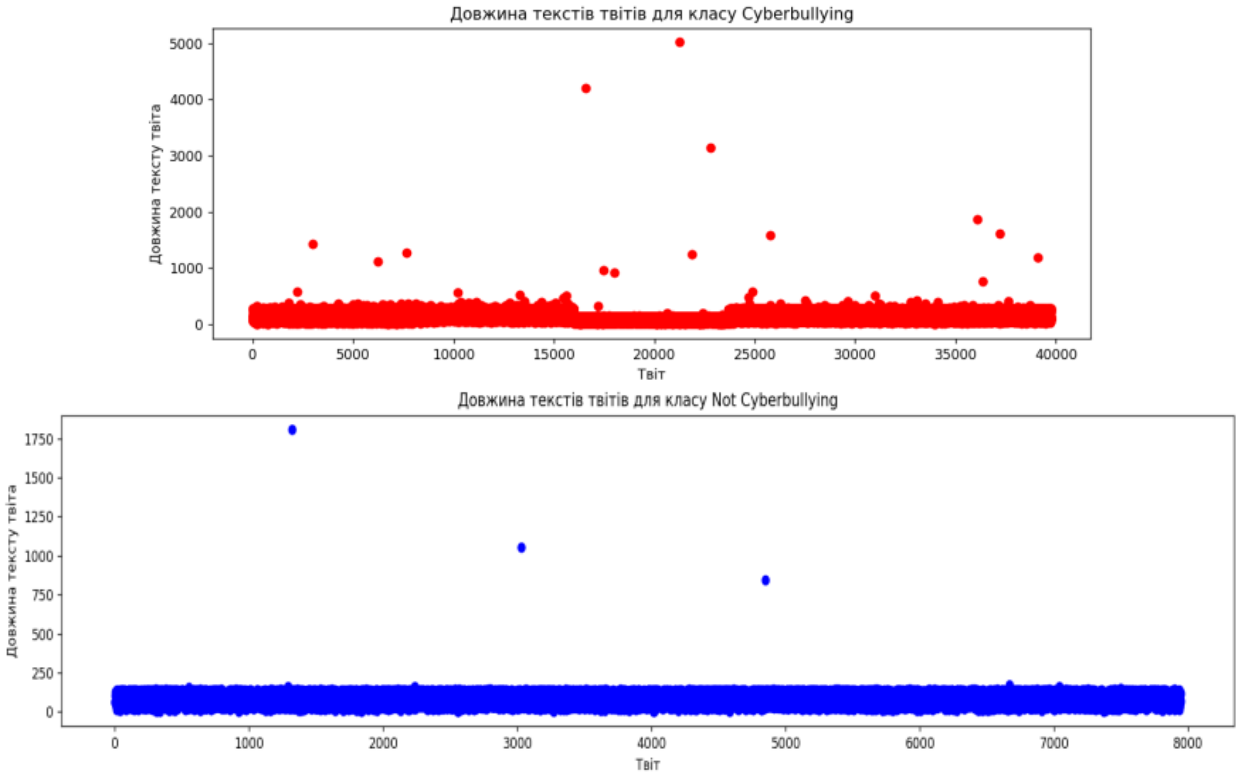
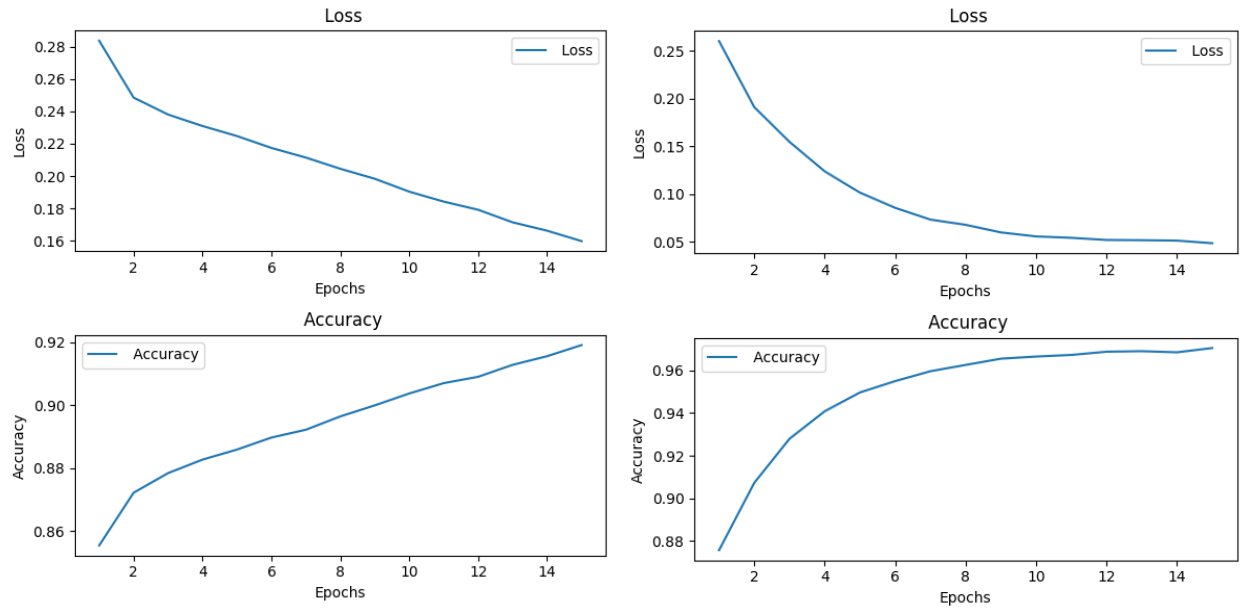


Рис. 5. Розподіл твітів за довжиною для класів «Cyberbullying» та «Not cyberbullying»

Також у подальшому варіант моделі з 25 епохами було збережено та досліджено із застосунку з графічним інтерфейсом. В цілому, такий розрив між навчальними та тестовими результатами пояснюється тим, що окрім простої фільтрації вибірки, не всі дані датасету мають коректну розмітку. Наприклад, зустрічаються твіти не лише англійською, а і португальською мовою: «*bullying é o oq passa o número 24 da caderneta de chamada !! #semgraça!, Not cyberbullying*». Хоча і твіт має розмітку як не кібербулінг, проте, його переклад «*Знущання - це те, про що говорить номер 24 у книзі викликів!!*» не на стільки однозначний. Програмно даний твіт класифікується як кібербулінг на 55%. Тому подальші дослідження будуть спрямовані на подальший аналіз датасету та його доповнення іншими даними, а також на покращення архітектури нейромережі з метою підвищення результатів класифікації.



а) **б)**

Рис. 6 Значення метрик втрат та точності: а) до фільтрації, б) після фільтрації

Розроблена інформаційна система має ряд обмежень, що пов’язані із даними для навчання нейромережі. Основне обмеження запропонованого підходу полягає у максимальній довжині вхідної послідовності, що складає 500 символів. Якщо вхідне повідомлення буде довшим, його величина буде

обрізана до 500 символів. Однак, враховуючи середню довжину твіта близько 200 символів, таке обмеження є обґрунтованим.

Висновки

У статті розглянуто поточний стан наукового напряму виявлення кібербулінгу, де на основі опрацьованого матеріалу запропоновано створення методу нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Запропонований підхід на основі об'єктно-орієнтованої моделі сприяє в створенні гнучких, адаптивних та легко розширюваних систем, а використанням хмарних сервісів для навчання нейромереж надає можливість не обмежуватись ресурсами виключно фізичного персонального комп'ютера, а задіяти спеціалізовані обчислювальні ресурси, такі як графічні процесори (GPU) або тензорні процесори (TPU), які значно прискорюють процес тренування нейронних мереж.

У якості даних дослідження було використано англомовний датасет Cyberbullying Classification», що налічував 39747 зразків, з яких було обрано 8000 зразків, що містять кібербулінг, та 7945 зразків без кібербулінгу.

Для дослідження ефективності запропонованого методу було створено програмну реалізацію, що написана мовою програмування Python у середовищі програмування PyCharm, а також ноутбук для виконання в хмарному сервісі «Google Colab» для навчання нейромережі.

Підхід має деякі обмеження, основне обмеження запропонованого підходу полягає у максимальній довжині вхідної послідовності, що складає 500 символів. Якщо вхідне повідомлення буде довшим, його величина буде обрізана до 500 символів. Однак, враховуючи середню довжину твіта близько 200 символів, таке обмеження є обґрунтованим. Для можливості покращення результату необхідно доповнити навчальний датасет та навести у ньому деякі обмеження по мінімальній та максимальній довжинам записів.

Література

1. Кібербулінг: як протистояти. URL: 2024. https://supportme.org.ua/needle-and-bullying/cyberbullying.

2. Кібербулінг. 2024. URL: https://www.unicef.org/ukraine/cyberbulling

3. Cyberbullying Classification. 2024. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.

4. European Court of Human Rights. 2024. URL: https://www.echr.coe.int/Pages/home.aspx?p=caselaw/otherpublications&c=#n 159309446013514343105 67_pointer

5. Detection of cyber bullying on social media using machine learning. 2024. URL: https://jespublication.com/upload/2022-V13I7091.pdf

6. Teoh H. T., Kasturi D. V. Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches. IEEE Access. 2023. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10122521.

7. Slobodzian V., Molchanova M., Kovalchuk O. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. ACIT. 2022. С. 400–405.

8. Молчанова М. О. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об\'єктно-орієтовану інформаційну систему / М. О. Молчанова, О. В. Мазурець, О. В. Собко // Вісник Хмельницького національного університету» серія: Технічні науки. – 2024. – №1 (331). – С. 101–106.

References

1. Kiberbulinh: yak protystoiaty. 2024. URL: https://supportme.org.ua/needle-and-bullying/cyberbullying.

2. Kiberbulinh. 2024. URL: https://www.unicef.org/ukraine/cyberbulling.

3. Cyberbullying Classification. 2024. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.

4. European Court of Human Rights. 2024. URL: https://www.echr.coe.int/Pages/home.aspx?p=caselaw/otherpublications&c=#n 159309446013514343105 67_pointer

5. Detection of cyber bullying on social media using machine learning. 2024. URL: https://jespublication.com/upload/2022-V13I7091.pdf

6. Teoh H. T., Kasturi D. V. Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches. IEEE Access. 2023. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10122521.

7. Slobodzian V., Molchanova M., Kovalchuk O. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. ACIT. 2022. S. 400–405.

8. Molchanova M. O. Alhorytm vyiavlennia abiuzyvnoho vmistu v ukrainomovnomu audiokontenti dlia implementatsii v ob\iektno-oriietovanu informatsiinu systemu / M. O. Molchanova, O. V. Mazurets, O. V. Sobko // Visnyk Khmelnytskoho natsionalnoho universytetu» seriia: Tekhnichni nauky. – 2024. – №1 (331). – S. 101–106.