Секція 1

ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ В ІНФОРМАЦІЙНОМУ СЕРЕДОВИЩІ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

Собко О. В.
Хмельницький національний університет
Науковий керівник: Бармак О. В.

Актуальність. Тема виявлення кіберзалякувань в інформаційному середовищі є актуальною у зв'язку з інтенсивним зростанням цифрового спілкування та постійним розширенням присутності користувачів в інтернет-просторі. Зокрема, розвиток соціальних мереж, онлайн-форумів та інших платформ для обміну інформацією створює сприятливі умови для анонімних кібератак та кіберманіпуляцій, які важко виявити та контролювати [1].

Метою даної роботи є розробка підходу до виявлення кіберзалякувань в інформаційному середовищі засобами машинного навчання.

Основні положення. Розроблено підхід до виявлення кіберзалякувань за текстовими зразками в інформаційному середовищі засобами машинного навчання. На початковому етапі вхідні дані, а саме текст для аналізу – проходить попередню обробку та векторизацію. Далі, використовуючи векторизатор та модель машинного навчання, спеціально навчену для розпізнавання кіберзалякувань, здійснюється оцінка ймовірності їх наявності в тексті. У результаті отримується висновок щодо присутності чи відсутності кіберзалякувань у аналізованому текстовому зразку. Розроблений підхід було апробовано шляхом створення програмного забезпечення, яке здатне автоматично визначати наявність кіберзалякувань у текстових повідомленнях. У даному програмному забезпечені для векторизації тексту та класифікації було використано модель BERT, яка зарекомендувала себе як ефективний інструмент для обробки природної мови. Модель була навчена на датасеті [2], який містив класи Age, Ethnicity, Gender, Religion, Other type of cyberbullying, Not cyberbullying процесі класифікації модель повертає значення ймовірності наявності кіберзалякування в тексті, а також надає результати щодо різних типів кіберзалякувань, що присутні в тексті, таких як вікові кіберзалякування, гендерні, релігійні, етнічні та інші. Модель продемонструвала високі показники якості класифікації, зокрема, значення метрик становили: макрометрик Accuracy 94%, Precision 93%, Recall 93%,

F1 Score 93%, що свідчить про високу точність та надійність моделі у виявленні кіберзалякувань.

Висновки. Розроблений підхід до виявлення кіберзалякувань в інформаційному середовищі засобами машинного навчання продемонстрував високу ефективність і точність, що робить його цінним інструментом у сфері кібербезпеки. Завдяки здатності моделі автоматично ідентифікувати образливий контент з високою ймовірністю правильного визначення, цей метод сприяє створенню безпечнішого інформаційного середовища. Запропоноване рішення допомагає запобігати поширенню кіберзалякувань, своєчасно виявляючи потенційно небезпечні повідомлення та знижуючи ризики негативного кібервпливу на користувачів, зокрема на дітей та молодь.

Список літератури

1. Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. CEUR Workshop Proceedings, 2024, Volume 3688, Page 16-28. DOI: https://doi.org/10.31110/COLINS/2024-3/002.

2. Cyberbullying Tweets. *Kaggle*. URL – https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets (дата звернення: 27.10.2024).

Відомості про авторів

Собко Олена Віталіївна, аспірантка кафедри комп'ютерних наук, Хмельницького національного університету, olenasobko.ua@gmail.com

Бармак Олександр Володимирович, завідувач кафедри комп'ютерних наук, Хмельницький національний університет, д.т.н., професор, alexander.barmak@gmail.com