ДОДАТОК А

Апробація результатів роботи

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

МАТЕРІАЛИ XXVIII МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ

«РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ»

16 – 18 квітня 2025 р.

Том 6

КОНФЕРЕНЦІЯ
«ІНФОРМАЦІЙНІ ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ»
INFORMATION INTELLIGENT SYSTEMS

Харків 2025

УДК 004.85

МОДЕРАЦІЯ ТЕКСТОВОГО КОНТЕНТУ З ВИКОРИСТАННЯМ КОМБІНОВАНОГО АНСАМБЛЕВОГО ПІДХОДУ НА ОСНОВІ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Керецман І.А.
e-mail: illia.keretsman@nure.ua
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

The object of the research is the process of automated text content moderation in digital environments.The aim of this work is to develop and evaluate the effectiveness of a combined approach to text content moderation using modern machine learning methods.The research methods include the analysis of existing text classification algorithms (Naive Bayes, SVM, Logistic Regression), the creation of an ensemble model with Gradient Boosting as a meta-model, and the evaluation of methods based on accuracy, precision, recall, and F1-score metrics.

Існуючі алгоритми модерації текстового контенту мають певні обмеження щодо точності виявлення токсичних повідомлень та їх класифікації. Класичні підходи, такі як наївний баєсівський класифікатор, логістична регресія та метод опорних векторів, показують високу продуктивність на простих наборах даних, але їхня ефективність суттєво знижується при обробці складних текстів, зокрема саркастичних або завуальованих образ. У той же час сучасні моделі на основі трансформерів (BERT, RoBERTa, GPT) демонструють значно кращі результати, проте їхнє навчання та використання потребують значних обчислювальних ресурсів.

Іншим ефективним підходом є застосування ансамблевих моделей, де комбінуються результати декількох алгоритмів для підвищення загальної продуктивності. Ансамблеві підходи дозволяють компенсувати слабкі сторони одних алгоритмів за рахунок сильних сторін інших, що підвищує загальну точність системи модерації [1].

Проведене дослідження дозволяє прийти до висновку, що застосування комбінованого ансамблевого підходу, який поєднує кілька класичних алгоритмів з мета-моделлю градієнтного бустингу (англ. Gradient Boosting). Градієнтний бустиг у якості мета моделі було обрано через його здатність:

– коригувати помилки попередніх моделей, поступово підвищуючи точність класифікації;

– ефективно працювати з дисбалансом даних, що є критичним для задач модерації тексту, де токсичний контент часто становить меншу частину загального обсягу тексту;

– забезпечувати високу продуктивність завдяки комбінації декількох слабких моделей у рамках одного ансамблю.

Такий підхід дозволяє використовувати сильні сторони різних алгоритмів та компенсувати їхні недоліки, підвищуючи загальну ефективність класифікації токсичного контенту.

Для порівняння ефективності було обрано три алгоритми класифікації тексту: наївний Баєсівський класифікатор, метод опорних векторів (*англ. SVM*), логістична регресія. Вибір цих алгоритмів зумовлений їхньою ефективністю у задачах текстової класифікації:

– наївний Баєсівський класифікатор є одним із найшвидших алгоритмів, який демонструє високу продуктивність при класифікації текстових даних, особливо коли важлива швидкість навчання та передбачення;

– метод опорних векторі обрано завдяки здатності створювати нелінійні межі між класами та ефективно працювати у високовимірних просторах;

– логістична регресія залишається одним із найбільш інтерпретованих алгоритмів, що добре працює у задачах бінарної класифікації, зокрема у випадках дисбалансу класів.

Для оцінки ефективності кожної з моделей та запропонованого ансамблю було використовуватися стандартні метрики машинного навчання: точність (*англ. accuracy*), влучність (*англ. precision*), повнота (*англ. recall*), F1-міра.

Тестування ефективності проводилось на Toxic Comment Dataset з платформи Kaggle [2]. Результати наведено в таблиці 1.

Таблиця 1 – Порівняння алгоритмів

Модель	Точність	Влучність	Повнота	F1
Наївний Баєс	95%	92%	51%	66%
SVM	88%	40%	57%	47%
Логістична регресія	96%	90%	61%	73%
Custom Ensemble (Gradient Bosting)	96%	84%	78%	76%

Комбінований ансамблевий підхід з використанням градієнтного бустингу показав найкращі результати серед усіх моделей. Базові моделі, такі як наївний Баєсівський класифікатор та логістична регресія, залишаються ефективними для швидкого розгортання, але їх ефективність значно поступається ансамблевим методам.

Список використаних джерел:

1. Text Classifiers in Machine Learning // IEEE Xplore. – 2024 – URL: https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guide (дата звернення 28.02.2025).

2. Jigsaw Toxic Comment Classification Challenge // Kaggle – 2018. – URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (дата звернення 28.02.2025).

ДОДАТОК Б

Звіт з результатами перевірки на унікальність тексту в базі ХНУРЕ

StrikePlagiarism.com

Дата звіту 6/13/2025
Дата редагування ---

Звіт не був оцінений

Звіт подібності

метадані

Назва організації
Kharkiv National University of Radio Electronics
Заголовок
2025_М_ПІ_ІПЗм-23-2_Керецман_І_А_скорочений
Автор
Керецман Ілля Андрійович
Науковий керівник / Експерт
Олена Олійник
підрозділ
каф. ПІ

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

8.19% 8.19% КП 1

1.26% 1.26% КЦ

25	6373	52031
Довжина фрази для коефіцієнта подібності 2	Кількість слів	Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв	ß	0
Інтервали	A→	0
Мікропробіли	·	0
Білі знаки	ß	0
Парафрази (SmartMarks)	a	30

Подібності за списком джерел

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз — Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	74 1.16 %
2	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	65 1.02 %
3	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	51 0.80 %

4	https://duikt.edu.ua/repozitorii/ai/2024/%D0%9E%D1%81%D1%82%D1%80%D0%B5%D0%BD%D1%81%D1%8C%D0%BA%D0%B8%D0%B9%20%D0%B4%D0%B8%D0%BF%D0%BB%D0%BE%D0%BC%20%D0%B1%D0%B0%D0%BA%D0%B0%D0%BB%D0%B0%D0%B2%D1%80%D0%B0%20%D0%A8%D0%86%D0%94-41.pdf	23 0.36 %
5	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	19 0.30 %
6	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	18 0.28 %
7	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	17 0.27 %
8	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	16 0.25 %
9	https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content	15 0.24 %
10	https://ena.lpnu.ua/bitstreams/da99cca3-c17f-41af-92eb-68f59c710495/download	15 0.24 %

з бази даних RefBooks (0.00 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з домашньої бази даних (0.50 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	2024_М_ШІ_СШІздм-22-1_Десятніченко_О_П_записка [redacted] 12/15/2024 Kharkiv National University of Radio Electronics (Kharkiv National University of Radio Electronics)	32 (4) 0.50 %

з програми обміну базами даних (0.85 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	Vasylchenko_Poiasniuvalna_zapyska_2024 [redacted] 12/13/2024 National Technical University "Kharkiv Polytechnic Institute" students papers (National Technical University "Kharkiv Polytechnic Institute" students papers)	13 (2) 0.20 %
2	Сокіл Р.О. ТРІМ-22.docx [redacted] 12/4/2018 National University "Lviv Politechnika" (NULP2)	12 (1) 0.19 %
3	Інтелектуальна система аналізу даних для комплексного управління клієнтським портфелем банку [redacted] 3/16/2025 National Technical University of Ukraine Igor Sikorskyi Kyiv Politech Institute (National Technical University of Ukraine Igor Sikorskyi Kyiv Politech Institute)	11 (2) 0.17 %
4	ФКНТ_2024_125_1_ШиленкоІС [redacted] 11/24/2024 Ukrainian national aviation university (ФКНТ Кафедра кібербезпеки)	11 (1) 0.17 %
5	2023_81220008_Pasternak_Rodion_Stepanovych_246086 [redacted] 10/26/2024 National University "Lviv Politechnika" (National University Lviv Politechnika)	7 (1) 0.11 %

з Інтернету (6.84 %)

ПОРЯДКОВИЙ НОМЕР	ДЖЕРЕЛО URL	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	**https://openarchive.nure.ua/server/api/core/bitstreams/e6565b39-9b83-497d-aa89-c8e9dd6fe289/content**	**315 (11) 4.94 %**

2	https://duikt.edu.ua/repozitorii/ipz/2025/%D0%9A%D0%B2%D0%B0%D0%BB%D1%96%D1%84%D1%96%D0%BA%D0%B0%D1%86%D1%96%D0%B9%D0%BD%D0%B0%20%D1%80%D0%BE%D0%B1%D0%BE%D1%82%D0%B0%20%D0%BC%D0%B0%D0%B3%D1%96%D1%81%D1%82%D1%80%D0%B0%20%D0%9B%D0%B5%D0%B2%D1%87%D0%B5%D0%BD%D0%BA%D0%B0%20%D0%9E.%D0%9E.%202025%20%D1%80%D1%96%D0%BA.pdf	26 (4) 0.41 %
3	https://duikt.edu.ua/repozitorii/ai/2024/%D0%9E%D1%81%D1%82%D1%80%D0%B5%D0%BD%D1%81%D1%8C%D0%BA%D0%B8%D0%B9%20%D0%B4%D0%B8%D0%BF%D0%BB%D0%BE%D0%BC%20%D0%B1%D0%B0%D0%BA%D0%B0%D0%BB%D0%B0%D0%B2%D1%80%D0%B0%20%D0%A8%D0%86%D0%94-41.pdf	23 (1) 0.36 %
4	https://ela.kpi.ua/bitstream/123456789/26892/1/Repiakh_magistr.pdf	21 (2) 0.33 %
5	https://ena.lpnu.ua/bitstreams/da99cca3-c17f-41af-92eb-68f59c710495/download	15 (1) 0.24 %
6	https://www.geeksforgeeks.org/k-means-vs-k-means-clustering-algorithm/	13 (2) 0.20 %
7	https://research.ijcaonline.org/volume89/number3/pxc3894222.pdf	12 (2) 0.19 %
8	http://pretrain.nlpedia.ai/	11 (1) 0.17 %

Список прийнятих фрагментів (немає прийнятих фрагментів)

ПОРЯДКОВИЙ НОМЕР	ЗМІСТ	КІЛЬКІСТЬ ОДНАКОВИХ СЛІВ (ФРАГМЕНТІВ)

1

ВСТУП

У сучасному цифровому середовищі текстовий контент є основним засобом комунікації, що активно використовується на різноманітних платформах, включаючи соціальні мережі, форуми, коментарі на сайтах новин та електронну пошту. Проте зростання обсягів текстового контенту супроводжується збільшенням кількості токсичних коментарів, образ, дезінформації та іншої шкідливої інформації. Це створює значний виклик для платформ та суспільства загалом. Автоматизація модерації тексту за допомогою методів машинного навчання стає критично важливою для забезпечення безпечного та інклюзивного цифрового простору.

Існуючі алгоритми модерації текстового контенту мають певні обмеження щодо точності виявлення токсичних повідомлень та їх класифікації. Класичні методи, такі **як наївний баєсівський класифікатор, логістична регресія та метод опорних векторів, показують високу продуктивність на простих наборах даних, але їхня ефективність суттєво знижується при обробці складних текстів, зокрема саркастичних або завуальованих образ. У той же час сучасні моделі на основі трансформерів (BERT, RoBERTa, GPT) демонструють значно кращі результати, проте їхнє навчання та використання потребують значних обчислювальних ресурсів.**

Дослідження ефективності нових методів машинного навчання, їх комбінування та вдосконалення є важливим етапом у вирішенні цієї проблеми. У рамках цієї роботи пропонується **застосування комбінованого** ансамблевого методу, який **поєднує кілька класичних алгоритмів з мета-моделлю градієнтного бустингу (англ. Gradient Boosting).** Такий метод дозволяє **використовувати сильні сторони різних алгоритмів та компенсувати їхні недоліки, підвищуючи загальну ефективність класифікації токсичного контенту.**

Метою роботи є розробка та порівняння ефективності комбінованого методу з базовими методами машинного навчання у контексті автоматизованої модерації

2

текстового контенту в цифрових середовищах. Для досягнення цієї мети необхідно вирішити наступні задачі:

− провести аналіз існуючих методів автоматизованої модерації текстового контенту;

− визначити сильні та слабкі сторони існуючих алгоритмів на основі метрик точності, повноти, влучності та F1-міри;

− запропонувати комбінований метод модерації, що поєднує декілька методів для підвищення ефективності за допомогою ансамблевої моделі;

− тестувати ефективність запропонованого методу на наборі даних Toxic

ДОДАТОК В

Слайди презентації

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Тема роботи

Дослідження методів машинного навчання для підвищення ефективності автоматизованої валідації та модерації текстового контенту в цифрових середовищах

Керецман Ілля Андрійович, ІПЗм-23-2
Науковий керівник: к.т.н., доц. Мельнікова Роксана Валеріївна

19 червня 2025

Дослідження

Актуальність та стан розвитку галузі

Обсяг текстового контенту в інтернеті стрімко зростає — водночас зростає й частка токсичних повідомлень. Автоматизація модерації тексту за допомогою методів машинного навчання стає критично важливою для забезпечення безпечного та інклюзивного цифрового простору.

Чітке визначення напряму дослідження

Оцінка ефективності методів машинного навчання у порівнянні з комбінованим методом на основі градієнтного бустингу у задачі модерації текстового контенту

Об’єкт дослідження

Об’єктом дослідження є методи машинного навчання у рамках автоматизованої модерації текстового контенту в цифрових середовищах.

Аналіз предметної області

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Reducing Bias and Improving Fairness in Machine Learning Models
- Jigsaw Toxic Comment Classification Challenge

Постановка задачі

Метою роботи є розробка та порівняння ефективності комбінованого методу з базовими методами машинного навчання у контексті автоматизованої модерації текстового контенту в цифрових середовищах. Для демонастрації роботи комбінованого методу розробити графічний інтерфейс користувача.

Очікується покращення ефективності базових методів шляхом використання їх у ансамблевому методі для навчання мета моделі.

Методологія

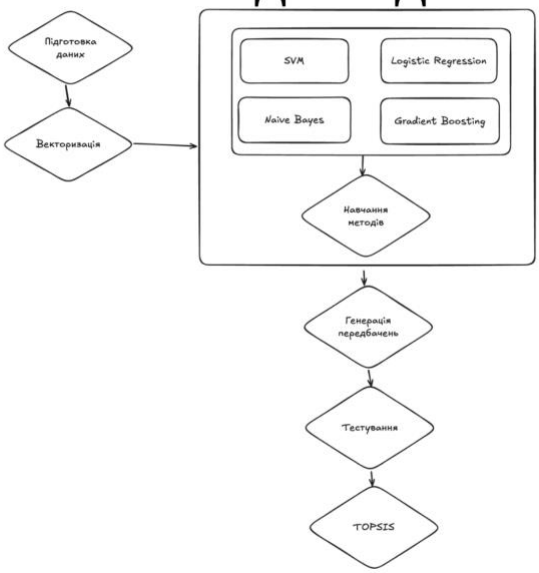
Опис методів дослідження

Методами дослідження є аналіз існуючих методів класифікації тексту (наївний Баєсівський класифікатор, метод опорних векторів, логістична регресія), створення ансамблевого методу з використанням градієнтного бустингу як мета-моделі, а також оцінка ефективності методів на основі метрик точності, влучності, повноти та F1-міри.

Інструменти та технології використані в роботі

У ході експерименту використані такі бібліотеки для мови програмування Python: TensorFlow, PyTorch, Hugging Face Transformers та Scikit-learn. Для навчання було використано датасет токсичних коментарів Toxic Comment Dataset з платформи Kaggle

Архітектура системи для проведення експериментального дослідження



Опис програмного забезпечення, що було використано у дослідженні

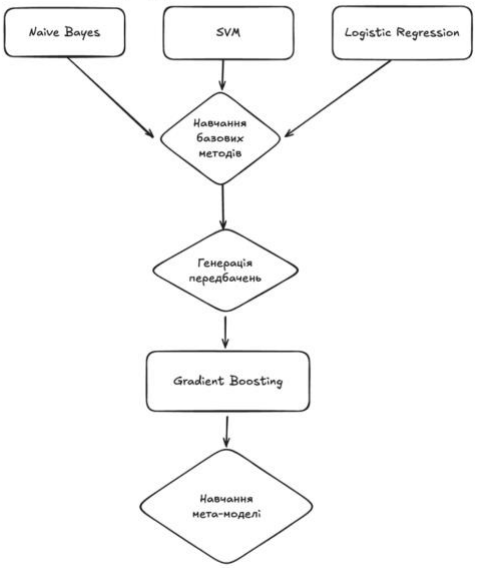
Опис процесу розробки

Робота включала написання комбінованого методу, де у якості мета-моделі є градієнтний бустинг. Далі було написано алгоритм порівняння трьох базових методів, а саме наївний Баєсівський класифікатор, метод опорних векторів, логістична регресія, та комбінованого підходу. Для відображення результатів порівняння та демонстрації практичного застосування розробленого комбінованого методу розроблений графічний інтерфейс користувача.

Вибрані мови програмування та фреймворки

Алгоритм для порівняння методів було написано на Python, з використанням бібліотек TensorFlow, PyTorch, Hugging Face Transformers та Scikit-learn. Фронтенд написано на React.

Зміст проведеного експерименту. Структура комбінованого методу



Зміст проведеного експерименту. Метрики

- точність
- влучність
- повнота
- F1-міра

9

Зміст проведеного експерименту. Вхідні дані

Toxic Comment Dataset Kaggle

Характеристики датасету:

- висока якість анотацій
- великий обсяг даних
- наявність різних типів токсичності

10

TOPSIS

Для визначення найефективнішого методу буде застосовано метод TOPSIS (*англ. Technique for Order Preference by Similarity to Ideal Solution*), який є одним з найпоширеніших підходів у багатокритеріальному аналізі. Вагові коефіцієнти для кожної метрики наведено нижче

Метрика	Тип	Вага
Точність	Максимізація	0.3
Влучність	Максимізація	0.15
Повнота	Максимізація	0.4
F1-міра	Максимізація	0.15

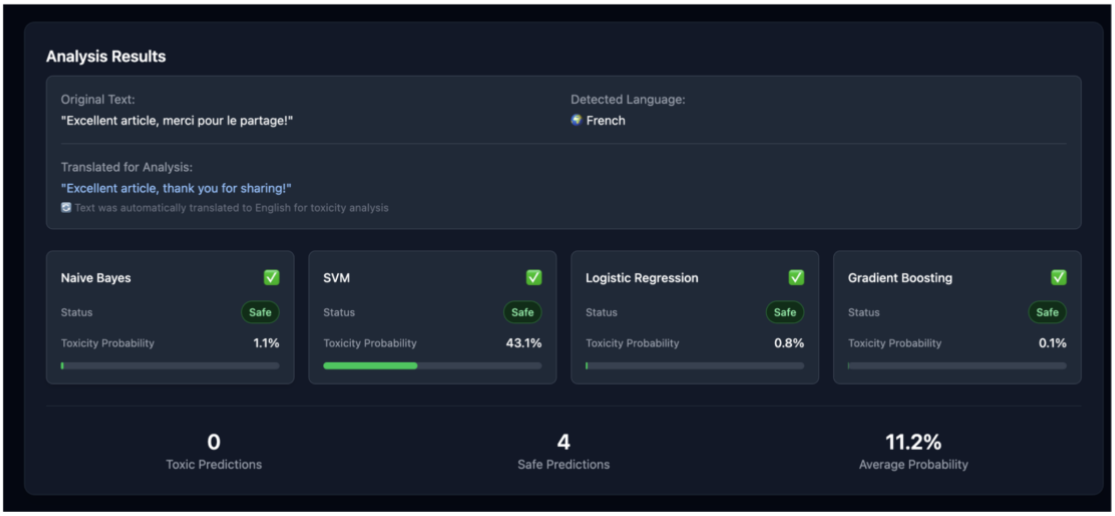
11

Результати експерименту

Метод	Точність	Влучність	Повнота	F1	Час навчання	Індекс TOPSIS
Наївний Баєс	95%	92%	51%	66%	0.01	**0.50**
SVM	88%	40%	57%	47%	18.63	**0.13**
Логістична регресія	96%	90%	61%	73%	0.23	**0.64**
Градієнтний бустинг	96%	84%	78%	76%	241.88	**0.90**

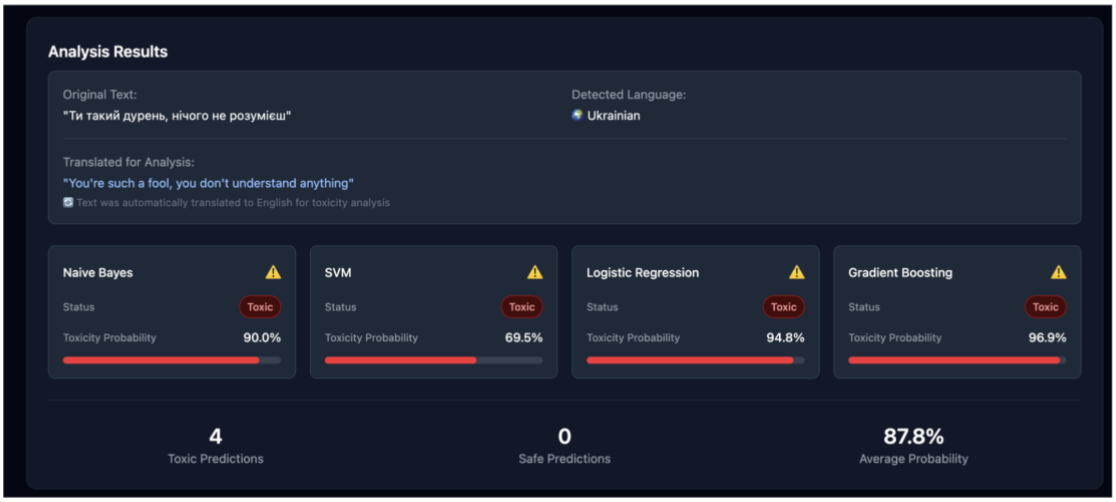
12

Демонстрація роботи комбінованого методу



13

Демонстрація роботи комбінованого методу

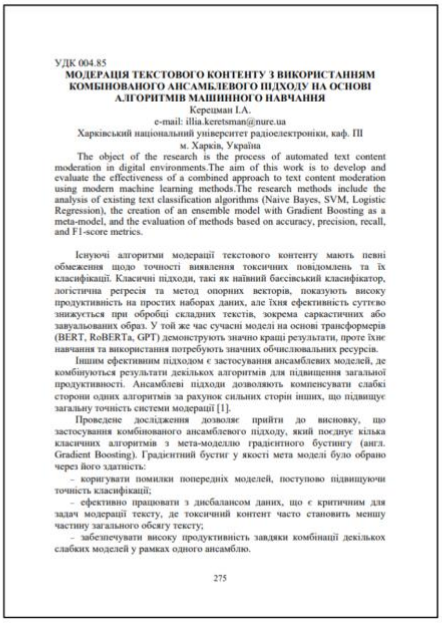


14

Аналіз отриманих результатів

Результати програмного експерименту показали, що комбінований ансамблевий метод продемонстрував найкращі результати, суттєво перевершивши базові методи за показниками повноти та F1-міри, що критично важливо для виявлення максимальної кількості токсичних коментарів та зменшення рівня false negative.

Публікація результатів



УДК 004.85

МОДЕРАЦІЯ ТЕКСТОВОГО КОНТЕНТУ З ВИКОРИСТАННЯМ КОМБІНОВАНОГО АНСАМБЛЕВОГО ПІДХОДУ НА ОСНОВІ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Керецман І.А.
e-mail: ilia.keretsman@nure.ua
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

The object of the research is the process of automated text content moderation in digital environments.The aim of this work is to develop and evaluate the effectiveness of a combined approach to text content moderation using modern machine learning methods.The research methods include the analysis of existing text classification algorithms (Naive Bayes, SVM, Logistic Regression), the creation of an ensemble model with Gradient Boosting as a meta-model, and the evaluation of methods based on accuracy, precision, recall, and F1-score metrics.

Існуючі алгоритми модерації текстового контенту мають певні обмеження щодо точності виявлення токсичних повідомлень та їх класифікації. Класичні підходи, такі як наївний баєсівський класифікатор, логістична регресія та метод опорних векторів, показують високу продуктивність на простих наборах даних, але їхня ефективність суттєво знижується при обробці складних текстів, зокрема саркастичних або завуальованих образ. У той же час сучасні моделі на основі трансформерів (BERT, RoBERTa, GPT) демонструють значно кращі результати, проте їхнє навчання та використання потребують значних обчислювальних ресурсів.

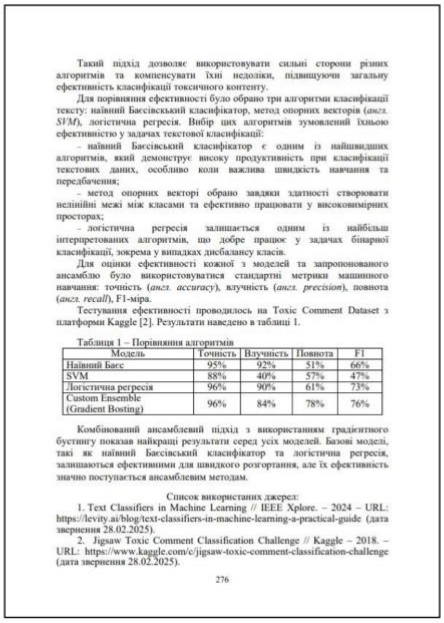
Іншим ефективним підходом є застосування ансамблевих моделей, де комбінуються результати декількох алгоритмів для підвищення загальної продуктивності. Ансамблеві підходи дозволяють компенсувати слабкі сторони одних алгоритмів за рахунок сильних сторін інших, що підвищує загальну точність системи модерації [1].

Проведене дослідження дозволяє прийти до висновку, що застосування комбінованого ансамблевого підходу, який поєднує кілька класичних алгоритмів з мета-моделлю градієнтного бустингу (англ. Gradient Boosting). Градієнтний бустинг у якості мета моделі було обрано через його здатність:

– коригувати помилки попередніх моделей, поступово підвищуючи точність класифікації;

– ефективно працювати з дисбалансом даних, що є критичним для задач модерації тексту, де токсичний контент часто становить меншу частину загального обсягу тексту;

– забезпечувати високу продуктивність завдяки комбінації декількох слабких моделей у рамках одного ансамблю.



Такий підхід дозволяє використовувати сильні сторони різних алгоритмів та компенсувати їхні недоліки, підвищуючи загальну ефективність класифікації токсичного контенту.

Для порівняння ефективності було обрано три алгоритми класифікації тексту: наївний Баєсівський класифікатор, метод опорних векторів (*англ. SVM*), логістична регресія. Вибір цих алгоритмів зумовлений їхньою ефективністю у задачах текстової класифікації:

– наївний Баєсівський класифікатор є одним із найшвидших алгоритмів, який демонструє високу продуктивність при класифікації текстових даних, особливо коли важлива швидкість навчання та передбачення;

– метод опорних векторі обрано завдяки здатності створювати нелінійні межі між класами та ефективно працювати у високовимірних просторах;

– логістична регресія залишається одним із найбільш інтерпретованих алгоритмів, що добре працює у задачах бінарної класифікації, зокрема у випадках дисбалансу класів.

Для оцінки ефективності кожної з моделей та запропонованого ансамблю було використовуватися стандартні метрики машинного навчання: точність (*англ. accuracy*), влучність (*англ. precision*), повнота (*англ. recall*), F1-міра.

Тестування ефективності проводилось на Toxic Comment Dataset з платформи Kaggle [2]. Результати наведено в таблиці 1.

Таблиця 1 – Порівняння алгоритмів

Модель	Точність	Влучність	Повнота	F1
Наївний Баєс	95%	92%	51%	66%
SVM	88%	40%	57%	47%
Логістична регресія	96%	90%	61%	73%
Custom Ensemble (Gradient Bosting)	96%	84%	78%	76%

Комбінований ансамблевий підхід з використанням градієнтного бустингу показав найкращі результати серед усіх моделей. Базові моделі, такі як наївний Баєсівський класифікатор та логістична регресія, залишаються ефективними для швидкого розгортання, але їх ефективність значно поступається ансамблевим методам.

Список використаних джерел:

1. Text Classifiers in Machine Learning // IEEE Xplore. – 2024 – URL: https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guide (дата звернення 28.02.2025).

2. Jigsaw Toxic Comment Classification Challenge // Kaggle – 2018. – URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (дата звернення 28.02.2025).

Підсумки

- проведено аналіз існуючих методів автоматизованої модерації текстового контенту;
- реалізовано комбінований метод, який об'єднує наївний Баєс, SVM та логістичну регресію з градієнтним бустингом як мета-моделлю;
- виконано програмну реалізацію системи модерації з графічним інтерфейсом користувача;
- проведено експерименти з використанням набору даних Toxic Comment Dataset;
- продемонстровано покращення метрик якості (F1, повнота, точність) порівняно з базовими методами;
- проміжні результати дослідження апробовані в рамках підготовки публікацій та захисту магістерської роботи;
- визначено перспективи розвитку системи: мультимовність, адаптація до інших типів платформ, інтеграція з існуючими рішеннями модерації.

Додаток Г

Експертний висновок результатів перевірки кваліфікаційної роботи

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-23-2 (група)

Керецман Ілля Андрійович
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
7.1.20	Заголовки структурних елементів звіту та заголовки розділів треба друкувати з абзацного відступу великими літерами напівжирним шрифтом без крапки в кінці. Дозволено їх розміщувати посередині рядка.	9, далі за текстом
	7.3 Нумерація сторінок звіту	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	

Експерт _____ (підпис) Вадим НЕЧВОЛОД (прізвище, ініціали)

13.06.2025