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 Баєсівскький 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).