Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

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

Виконав:
здобувач 2 року навчання
групи ІПЗм-23-2

Ілля КЕРЕЦМАН
(Власне ім'я, ПРІЗВИЩЕ)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)
Тип програми освітньо-наукова

Керівник доц. Роксана МЕЛЬНІКОВА
(посада, Власне ім'я, ПРІЗВИЩЕ)

Допускається до захисту
Зав. кафедри _____ Кирило СМЕЛЯКОВ
(підпис) (Власне ім'я, ПРІЗВИЩЕ)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Керецману Іллі Андрійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів машинного для підвищення ефективності автоматизованої валідації та модерації текстового контенту в цифрових середовищах»

Затверджена наказом по університету від 15.04. 2025р. № 290 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 19.06.2025

3. Вихідні дані до роботи: огляд й аналіз літературних, наукових джерел, теоретичне дослідження, проведення експерименту для порівняння ефективності методів машинного навчання для модерації текстового контенту, розробка комбінованого методу.

4. Перелік питань, що потрібно опрацювати в роботі
вступ, аналіз предметної галузі, огляд й аналіз літературних, наукових джерел, постановка задачі, теоретичне дослідження, проведення експерименту, практичне застосування, висновки.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
	Отримання завдання	17.04.2025	*виконано*
	Аналіз предметної галузі і постановка задачі	17.04.2025	*виконано*
	Огляд й аналіз літературних, наукових джерел	17.04.2025	*виконано*
	Теоретичне дослідження	05.05.2025	*виконано*
	Проведення експерименту	10.05.2025	*виконано*
	Підготовка до апробації результатів дослідження. Публікація матеріалів	15.05.2025	*виконано*
	Підготовка пояснювальної записки	29.05.2025	*виконано*
	Підготовка презентації та доповіді	02.06.2025	*виконано*
	Перевірка на плагіат	12.06.2025	*виконано*
	Нормоконтроль	13.06.2025	*виконано*
	Рецензування	14.06.2025	*виконано*
	Попередній захист	17.06.2025	*виконано*
	Занесення диплома в електронний архів	17.06.2025	*виконано*
	Допуск до захисту у зав. кафедри	17.06.2025	*виконано*

Дата видачі завдання 16 квітня 2025р.

Студент _____ Ілля КЕРЕЦМАН
(підпис)

Керівник роботи _____ доц. Роксана МЕЛЬНІКОВА
(підпис) (посада, Власне ім’я, ПРІЗВИЩЕ)

РЕФЕРАТ / ABSTRACT

Робота містить: 65 с., 4 табл., 17 джерел, 12 формул, 3 рисунка.

АНАЛІЗ, АНСАМБЛЕВІ МОДЕЛІ, ГРАДІЄНТНИЙ БУСТИНГ, ЛОГІСТИЧНА РЕГРЕСІЯ, МАШИННЕ НАВЧАННЯ, МЕТОД ОПОРНИХ ВЕКТОРІВ ,МОДЕРАЦІЯ, ТОКСИЧНИЙ КОНТЕНТ, BERT, NLP, SVM, TF-IDF.

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

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

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

У результаті роботи було створено комбінований метод, а також програмна реалізація методу на мові програмування Python, протестовано його на наборі даних Toxic Comment Dataset та отримано підвищення ефективності у порівнянні з базовими методами машинного навчання.

ANALYSIS, ENSEMBLE MODELS, GRADIENT BOOSTING, LOGISTIC REGRESSION, MACHINE LEARNING, SUPPORT VECTOR MACHINES, MODERATION, TOXIC CONTENT, BERT, NLP, SVM, TF-IDF.

The object of the research are machine learning methods within the framework of automated moderation of textual content in digital environments.

The aim of this work is to develop and evaluate the effectiveness of a combined method 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.

As a result of the work, a combined method was developed, and also program implementation of method written in Python programming language, tested on the Toxic Comment Dataset, and demonstrated increased efficiency compared to baseline machine learning methods.

Завідувачу кафедри

ПІ

(скорочена назва кафедри)

проф. Кирилу СМЕЛЯКОВУ

(вчене звання, сласне ім’я, прізвище)

ЗАЯВА

щодо самостійності виконання кваліфікаційної роботи та можливості її публікації (та/або публікації анотації кваліфікаційної роботи) в електронному архіві відкритого доступу ElAr KhNURE

Я, Керецман Ілля Андрійович

(прізвище, ім’я, по батькові)

здобувач вищої освіти на другому (магістерському) рівні вищої освіти академічної групи ІПЗм-23-2

кафедра програмної інженерії ,

(повна назва кафедри)

заявляю: моя кваліфікаційна робота на тему «Дослідження методів машинного навчання для підвищення ефективності автоматизованої валідації та модерації текстового контенту в цифрових середовищах», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в репозиторії "ElArKhNURE". Погоджуюся з авторським договором, відповідно до Положення про репозиторій ХНУРЕ "ElArKhNURE". Всі запозичення з друкованих та електронних джерел мають відповідні посилання. Я ознайомлений (а) з вимогами академічної доброчесності, згідно з якими виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

Дата 09.06.2025 Підпис

ЗМІСТ

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

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 Comment Dataset з платформи Kaggle;
- порівняти результати з існуючими рішеннями та зробити висновки щодо доцільності використання комбінованого методу.

Об’єктом дослідження є процеси автоматизованої модерації текстового контенту в цифрових середовищах, а предметом – методи машинного навчання для валідації та модерації текстового контенту, зокрема ансамблеві методи до класифікації токсичних коментарів.

У ході роботи будуть використані наступні методи дослідження:

- аналіз літератури та існуючих рішень для вивчення поточного стану проблеми та існуючих методі до її вирішення;
- розробка алгоритму для створення комбінованого ансамблевого методу до модерації текстового контенту з використанням базових моделей (наївний Баєсівський класифікатор, метод опорних векторів, логістична регресія) та мета-моделі градієнтний бустинг;
- аналіз даних для оцінки ефективності методів на основі метрик точності, повноти, влучності та F1-міри;
- аналіз для оцінки ефективності запропонованого методу у порівнянні з існуючими методами.

Результати дослідження можуть бути використані в практичній діяльності компаній, що займаються розробкою систем модерації контенту, а також в академічних дослідженнях для подальшого розвитку цієї галузі.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

У цифровому середовищі, де кількість текстової інформації зростає експоненціально, актуальність ефективних рішень для автоматизованої модерації стає критичною. Збільшення обсягів контенту на платформах соціальних мереж, форумах та сайтах новин призводить до необхідності швидкого та точного реагування на порушення правил спільнот. Метою цього розділу є узагальнення та систематизація існуючих методів, обмежень та перспективних напрямків розвитку в галузі автоматизованої модерації контенту.

1.1 Тенденції та перспективи

Останнім часом спостерігається активний розвиток мультимодальних моделей, які враховують не лише текст, але й зображення, аудіо та відеоконтент. Цей підхід дозволяє створювати комплексні системи модерації, які здатні аналізувати кілька форматів даних одночасно, забезпечуючи виявлення токсичних елементів навіть у складних комбінаціях тексту та візуальних матеріалів.

Компанії Google, Facebook та Microsoft активно впроваджують мультимодальні системи модерації, які поєднують аналіз тексту з обробкою зображень і відео. Наприклад, Google інтегрує аналіз текстів і відео для модерації контенту на платформі YouTube, що дозволяє ефективніше виявляти порушення політик платформи [1].

Додатково розвивається напрямок моделей самонавчання (*англ. self-supervised learning*), які дозволяють алгоритмам навчатися на великих обсягах неанотованих даних. Це дає можливість зменшити потребу у ручному маркуванні, що суттєво спрощує адаптацію систем до нових мовних шаблонів і текстових трендів. Такі моделі демонструють високу ефективність у боротьбі з завуальованою токсичністю та можуть оперативно оновлювати свої знання без людського втручання [2].

Важливою складовою розвитку також є впровадження передавального навчання, яке дозволяє переносити знання з одних доменів на інші, що підвищує

універсальність та адаптивність моделей до різних платформ і мовних середовищ. У поєднанні з мультимодальними підходами це сприяє створенню більш стійких систем модерації, здатних ефективно обробляти змішаний контент та реагувати на нові форми онлайн-цькування.

1.2 Огляд існуючих підходів

У сучасній практиці для модерації текстового контенту активно застосовуються різні методи машинного навчання, включаючи класичні алгоритми (наївний Баєсівський класифікатор, метод опорних векторів, логістична регресія) та більш сучасні підходи, такі як глибокі нейронні мережі (BERT, DistilBERT, RoBERTa, GPT). Ці моделі використовуються для аналізу текстів, класифікації та виявлення образливих або токсичних висловлювань. Моделі на базі трансформерів демонструють високу ефективність у виявленні токсичного контенту завдяки здатності до контекстуального аналізу тексту та врахування взаємозв'язків між словами в реченні [3].

Одним із найпопулярніших підходів є використання моделей BERT, які дозволяють аналізувати текст на рівні семантики та виявляти токсичні висловлювання. Наприклад, модель BERT була використана у відомому конкурсі Toxic Comment Classification Challenge на платформі Kaggle, де учасники розробляли рішення для автоматизації модерації текстового контенту. У рамках цього конкурсу моделі навчалися на великому наборі даних, що складався з коментарів користувачів із позначками токсичності [4].

Іншим ефективним підходом є застосування ансамблевих методів, де комбінуються результати декількох алгоритмів для підвищення загальної продуктивності. Наприклад, комбінація методу опорних векторів з нейронними мережами або градієнтний бустинг часто демонструють високі результати у виявленні небажаного контенту. Ансамблеві методи дозволяють компенсувати слабкі сторони одних алгоритмів за рахунок сильних сторін інших, що підвищує загальну точність системи модерації [5].

1.3 Обмеження існуючих рішень

Попри високу ефективність сучасних моделей, існує низка обмежень, які заважають повністю автоматизувати процес модерації. Основною проблемою є низька точність при обробці непрямої або завуальованої мови. Наприклад, токсичні висловлювання можуть бути представлені в саркастичній або іронічній формі, що ускладнює їх автоматичне виявлення. У багатьох випадках токсичні висловлювання маскуються, використовуючи евфемізми або альтернативні написання слів [6].

Сучасні алгоритми також часто демонструють упередженість щодо певних соціальних груп, що призводить до великої кількості хибнопозитивних результатів. Ця проблема виникає через нерівномірність розподілу даних у навчальних вибірках або через наявність упередженості в самому наборі даних. Для подолання цієї проблеми дослідники розробляють нові методи зменшення упередженості шляхом збільшення репрезентативності навчальних вибірок та застосування технік розширення даних, що дозволяють збільшити різноманітність текстів у наборі даних [7].

Традиційні підходи до семантичного аналізу часто не враховують специфіку функціонального стилю тексту, що призводить до втрати частини смислового навантаження та зниження точності при класифікації складних мовних конструкцій. Окремі дослідження також акцентують на цьому недоліку та пропонують узагальнені алгоритми для покращення визначення семантичної подібності між текстами різних стилістичних типів. [8]

1.4 Масштаб проблеми

Масштаб проблеми токсичного контенту постійно зростає, адже з кожним роком кількість користувачів інтернету збільшується, що спричиняє ріст обсягів контенту. Онлайн-ресурси, такі як соціальні мережі та форуми, щодня генерують мільйони коментарів, значна частина з яких може містити образливий або токсичний зміст. За даними Pew Research Center, понад 41% користувачів інтернету

зіштовхуються з онлайн-цькуванням або образами, що свідчить про масштабність проблеми та потребу у її вирішенні [9].

Крім того, дослідження, проведене компанією Microsoft у рамках проекту Digital Civility Index, виявило, що 70% молодих користувачів інтернету регулярно стикаються з токсичним контентом у соціальних мережах. Це вказує на серйозність проблеми та необхідність розробки більш ефективних інструментів для боротьби з онлайн-цькуванням та іншими формами агресивної поведінки [10].

1.5 Визначення рівня інноваційності

Розробка нових підходів до автоматизованої модерації текстового контенту потребує впровадження інноваційних методів, що дозволяють адаптуватися до змін у мовних шаблонах та типах токсичних повідомлень. Одним із ключових аспектів є використання самонавчальних алгоритмів, які можуть динамічно оновлювати свої моделі на основі нових даних без необхідності ручного втручання. Такі методи включають самокероване навчання, передавальне навчання та навчання з підкріпленням, що дає змогу підвищити ефективність модерації та зменшити кількість помилкових спрацьовувань.

Важливим напрямом є також інтеграція мультимодальних моделей, які поєднують текстовий, візуальний та аудіоконтент для створення комплексної системи оцінки. Це дозволяє краще аналізувати контекст повідомлення, знижуючи кількість помилкових рішень при обробці саркастичних або прихованих токсичних висловлювань [11].

Загальний рівень інноваційності також визначається здатністю моделей до швидкої адаптації до нових типів даних, що є критично важливим у середовищах, де мова та форми спілкування постійно змінюються.

2 ОГЛЯД ТА АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

2.1 Огляд основних джерел

Під час добору джерел особлива увага приділялася таким критеріям, як авторитетність публікацій (IEEE, ACM, arXiv), актуальність (джерела за останні 5 років), а також об'єктивність і достовірність даних.

2.1.1 Теоретична база для моделювання трансформерів

Одним із ключових джерел є "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [3]. У цій статті детально описана методологія створення моделі BERT, яка стала проривом у сфері обробки природної мови. Основною особливістю моделі є її здатність враховувати контекст слова як зліва, так і справа, що суттєво підвищує точність аналізу тексту.

Ця робота стала основою для багатьох подальших досліджень у сфері NLP, включаючи автоматизовану модерацію текстів. Зокрема, можливість обробляти складні речення та розпізнавати завуальовані образи в тексті робить BERT одним із найефективніших інструментів для вирішення завдань класифікації токсичного контенту.

2.1.2 Практичні результати застосування модераційних моделей

Досвід конкурсу "Jigsaw Toxic Comment Classification Challenge" [4] є цінним джерелом для аналізу ефективності різних підходів у класифікації токсичних коментарів. У рамках цього конкурсу використовувалися великі набори даних із вручну анотованими прикладами токсичних висловлювань. Це дозволило учасникам розробляти та тестувати інноваційні моделі, включаючи трансформери.

Результати конкурсу підтвердили ефективність використання моделей BERT і RoBERTa, а також підкреслили важливість ансамблевих методів для підвищення точності класифікації. Окрім того, значний обсяг даних дозволив виявити переваги використання великих анотованих корпусів для навчання моделей, що має вирішальне значення для підвищення точності у реальних умовах.

2.1.3 Зменшення упередженості в моделюванні

Стаття "Reducing Bias and Improving Fairness in Machine Learning Models" [7] акцентує увагу на важливості зменшення упередженості у моделях машинного навчання. У роботі розглядаються ситуації, коли нерівномірний розподіл даних у навчальних вибірках призводить до дискримінації окремих соціальних груп. Це особливо важливо для систем модерації контенту, де високий рівень хибнопозитивних результатів може призводити до неправомірного блокування користувачів.

У статті запропоновано кілька методів для вирішення цієї проблеми, включаючи:

- збільшення репрезентативності вибірок шляхом підбору даних із різних джерел;
- застосування технік data augmentation для збільшення різноманітності навчальних даних;
- оптимізацію архітектури моделей, що дозволяє краще враховувати контекст і зменшувати рівень упередженості.

Це джерело є фундаментальним для розробки більш етичних і справедливих моделей автоматизованої модерації, що особливо важливо для побудови довіри користувачів до таких систем.

2.1.4 Автоматизація модерації в групових чатах

Практичне застосування алгоритмів модерації в групових чатах, таких як Telegram, детально розглянуто у статті "Використання алгоритмів машинного навчання для автоматизації процесу модерації контенту в групових чатах месенджерів" [12]. У роботі підкреслюється, що зростання обсягів повідомлень і різноманітність контенту створюють серйозні виклики для ручної модерації.

Автори статті аналізують ефективність використання класичних моделей (SVM, логістична регресія) та сучасних трансформерів для автоматизації процесу

модерації. Результати показують, що комбінований метод із використанням ансамблю може значно підвищити точність фільтрації небажаного контенту.

2.1.5 Огляд ансамблевих методів

Ансамблеві методи машинного навчання дозволяють підвищити точність класифікації шляхом поєднання декількох базових моделей. Найпоширенішими серед них є беггінг (*англ. bagging*) і бустинг (*англ. boosting*), кожен з яких має свої особливості, переваги та сфери застосування.

Беггінг працює за принципом паралельного навчання кількох моделей на випадкових підмножинах навчального набору з подальшим агрегуванням результатів. Це дозволяє зменшити дисперсію моделі та уникнути надмірного навчання. Беггінг є ефективним у випадках, коли базові класифікатори мають високу варіативність і нестабільність.

Натомість бустинг навчає базові моделі послідовно, причому кожна наступна модель фокусується на тих прикладах, які були неправильно класифіковані попередніми. Бустинг є ефективною стратегією побудови ансамблевого класифікатора, оскільки на кожному етапі він враховує помилки попередніх моделей і адаптивно навчається на важких прикладах [13].

Крім того, градієнтний бустинг дозволяє використовувати довільні диференційовані функції втрат, що розширює його застосування для різних типів задач. На відміну від беггінгу, що лише знижує дисперсію, бустинг зменшує як дисперсію, так і зміщення – а отже, забезпечує більш збалансоване навчання [14].

У контексті задач модерації текстового контенту, де присутній дисбаланс класів і необхідність виявлення неочевидної токсичності, бустинг у теорії демонструє кращі результати. Завдяки здатності фокусуватись на складних прикладах, цей метод краще виявляє рідкісні класи, що критично важливо для побудови ефективної системи автоматичної модерації.

2.1.6 Методи виявлення синтетичних текстів

Окремим напрямом досліджень у сфері класифікації текстового контенту є виявлення синтетичних або автоматично згенерованих текстів. З розвитком генеративних мовних моделей зросла актуальність задачі ідентифікації контенту, створеного нейромережами, оскільки такі тексти можуть бути використані для масового розповсюдження токсичних або маніпулятивних повідомлень.

У роботі [15] було проведено порівняльний аналіз сучасних підходів до виявлення синтетичних текстів, серед яких — застосування логістичної регресії, методів на основі трансформерів (зокрема RoBERTa), а також zero-shot класифікації. Дослідження показало, що використання трансформерних моделей забезпечує високу якість розпізнавання згенерованого контенту навіть без додаткового навчання на спеціалізованих наборах даних.

Варто зазначити, що подібні підходи можуть бути адаптовані і для задач автоматизованої модерації текстового контенту, зокрема для виявлення завуальованої або слабо вираженої токсичності у текстах, створених за допомогою генеративних моделей. Це відкриває перспективи подальшого поєднання традиційних методів класифікації із сучасними підходами до детекції синтетичного контенту в рамках комплексних систем модерації.

2.2 Оцінка актуальності та новизни

Аналіз джерел показав, що автоматизація модерації контенту залишається важливим напрямком досліджень, оскільки кількість токсичних повідомлень у цифрових середовищах зростає. Особливу увагу приділяють методам зменшення упередженості, що свідчить про зростання актуальності цих тем у науковій спільноті.

3 ПОСТАНОВКА ЗАДАЧІ

3.1 Визначення тематики та кінцевих результатів

Тематика цього проєкту спрямована на дослідження автоматизованої модерації текстового контенту із застосуванням сучасних моделей машинного навчання. У ході роботи потрібно розробити комбінований ансамблевий метод, який дозволить підвищити ефективність процесу модерації, мінімізуючи кількість помилок у виявленні токсичного контенту. Особливу увагу буде приділено зменшенню кількості хибних спрацьовувань та покращенню здатності методу виявляти завуальовані образи.

3.2 Обґрунтування вибору методів дослідження

У ході дослідження передбачається здійснення аналізу коментарів для оцінки існуючих рішень та виділення їхніх переваг і недоліків. Важливо буде порівняти продуктивність різних методів класифікації, зокрема наївного Баєсівського класифікатора (*англ. Naive Bayes*), методу опорних векторів (*англ. SVM, support vectore machine*) та логістичної регресії (*англ. Logistic Regression*).

Вибір цих методів зумовлений їхньою ефективністю у задачах текстової класифікації:

- наївний Баєсівський класифікатор є одним із найшвидших методів, який демонструє високу продуктивність при класифікації текстових даних, особливо коли важлива швидкість навчання та передбачення;
- SVM обрано завдяки здатності створювати нелінійні межі між класами та ефективно працювати у високовимірних просторах;
- логістична регресія залишається одним із найбільш інтерпретованих методів, що добре працює у задачах бінарної класифікації, зокрема у випадках дисбалансу класів.

Окрім класичних методів, у цьому дослідженні також буде використано ансамблевий метод із застосуванням градієнтного бустингу у якості мета-моделі. Градієнтний бустиг у якості мета моделі було обрано через його здатність:

- коригувати помилки попередніх методів, 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 методів у рамках одного ансамблю.

Градієнтний бустинг використовується як мета-модель для ансамблю з наївним Баєсівським класифікатором, SVM та логістичною регресією, дозволяючи компенсувати слабкі сторони базових методів та підвищити загальну продуктивність системи.

Такий комбінований метод забезпечує більш точну та стабільну модерацію текстового контенту, дозволяючи системі краще справлятися з варіативними текстами, включаючи саркастичні або завуальовані токсичні висловлювання.

Для оцінки ефективності кожного з методів та запропонованого ансамблю будуть використовуватися стандартні метрики машинного навчання:

- точність (*англ. accuracy*) – показує частку правильно класифікованих прикладів;
- влучність (*англ. precision*) – вимірює, наскільки модель точна у передбаченні токсичних коментарів;
- повнота (*англ. recall*) – демонструє здатність моделі виявляти всі токсичні приклади;
- F1-міра – дозволяє збалансувати точність та повноту, що особливо важливо у випадках дисбалансу класів.

Використання цих метрик дозволить комплексно оцінити якість методів, враховуючи як точність передбачень, так і їхню здатність виявляти максимальну кількість токсичних коментарів із мінімальною кількістю хибнопозитивних результатів.

3.3 Обмеження дослідження

Одним із ключових викликів, які постають перед дослідниками в галузі автоматизованої модерації текстового контенту, є мовні бар’єри та специфіка обробки текстів різними мовами. Більшість існуючих досліджень зосереджуються на англомовному контенті, що зумовлено доступністю великих анотованих корпусів англійської мови. Проте, ефективність методів, навчання яких базується виключно на англомовних даних, часто виявляється значно нижчою при застосуванні до текстів іншими мовами. Це пов'язано з лінгвістичними особливостями, синтаксичними та морфологічними відмінностями мов, що створює додаткові труднощі у перенесенні моделей на нові мови без повторного навчання.

Окрім мовного бар’єру, важливим обмеженням є залежність від обсягу та якості анотованих даних. Більшість сучасних методів глибокого навчання демонструють високі результати лише за наявності великих та добре структурованих наборів даних. У випадку з текстовою модерацією, анотація даних є трудомістким і дорогим процесом, що вимагає залучення експертів для забезпечення якісного маркування. Навіть у межах одного мовного середовища токсичність може бути представлена різними формами – від явних образ до саркастичних і прихованих висловлювань, що ускладнює процес анотації та знижує ефективність моделей.

Ще одним важливим аспектом є потреба у значних обчислювальних потужностях для навчання та використання великих трансформерних моделей, таких як BERT, RoBERTa або GPT. Процес навчання моделей з великою кількістю параметрів може займати години або навіть дні, що ускладнює їхню адаптацію та оптимізацію. Використання хмарних обчислень або оренда GPU-серверів може значно збільшити витрати на проєкт. У випадку обмежених обчислювальних ресурсів дослідники змушені звертатися до менш складних моделей, що може призвести до зниження загальної продуктивності та якості класифікації.

Додатковим викликом є упередженість моделей, яка виникає внаслідок нерівномірного розподілу класів у навчальних вибірках. Наприклад, токсичні коментарі часто становлять менше ніж 10% від загального обсягу тексту, що може призвести до переважання нетоксичних коментарів у результатах класифікації. Це може спричинити значну кількість хибнонегативних спрацювань, коли токсичні коментарі залишаються непоміченими моделлю.

Таким чином, основні обмеження дослідження включають мовну специфіку, обмеженість анотованих даних, потребу у високих обчислювальних потужностях та можливі упередження моделей, що потребують додаткових методів для їхнього усунення.

3.4 Програмна імплементація експерименту

Метою дослідження є розробка програмної імплементації експерименту за допомогою мови програмування Python. Програма має представляти собою реалізацію комбінованого методу та порівняння його з існуючими методами машинного навчання (наївний Баєсівський класифікатор, SVN, логістична регресія). Для відображення результатів порівняння та демонстрації практичного застосування розробленого комбінованого методу має буте розроблений графічний інтерфейс. Для часткового вирішення мовної проблеми робота має адаптувати застосування комбінованого методу для іноземних мов за допомогою машинного перекладу.

У якості даних для навчання і тестування методів обрано Toxic Comment Dataset з платформи Kaggle[4]. Цей датасет є одним із найбільших і найвідоміших у сфері текстової модерації та містить понад 159 000 коментарів з анотаціями, які вказують на рівень токсичності.

Ключовими перевагами цього датасету є:

- висока якість анотацій – коментарі були марковані вручну, що підвищує точність навчання моделей;
- великий обсяг даних – значна кількість текстових прикладів дозволяє навчати складні моделі без ризику перенавчання;

- наявність різних типів токсичності – датасет містить мітки, що вказують на різні категорії токсичного контенту (образи, погрози, ненормативна лексика тощо), що дає змогу створювати більш універсальні моделі модерації.

Для забезпечення ефективної роботи з даними також потрібне використання бібліотек для обробки тексту. У ході експерименту будуть використані такі бібліотеки для мови програмування Python: TensorFlow, PyTorch, Hugging Face Transformers та Scikit-learn. Ці інструменти дозволяють працювати з текстовими даними на різних рівнях, включаючи попередню обробку, векторизацію та створення моделей.

Завдяки доступу до всіх необхідних ресурсів можна забезпечити ефективне навчання моделей та досягти високих результатів у задачах модерації текстового контенту.

4 ТЕОРЕТИЧНЕ ДОСЛІДЖЕННЯ

4.1 Огляд основних методів

У цьому розділі буде розглянуто три основні методи до класифікації тексту: наївний Баєсівський класифікатор, метод опорних векторів та логістична регресія.

Вибір саме цих методів зумовлений їхньою широкою застосовністю в задачах обробки тексту та доведеною ефективністю при класифікації коротких текстових фрагментів, таких як коментарі та дописи у соціальних мережах. Кожен із цих методів використовує різні підходи до класифікації, що дозволяє отримати більш комплексне розуміння даних та підвищити точність кінцевого методу.

4.1.1 Наївний Баєсівський класифікатор

Наївний Баєсівський класифікатор є одним із найпростіших і найшвидших методів для задач класифікації тексту. Його популярність обумовлена високою швидкістю обробки даних і ефективністю навіть на великих обсягах текстової інформації. Метод базується на застосуванні теореми Баєса, яка дозволяє обчислювати ймовірність належності тексту до певного класу на основі попередніх спостережень.

Однією з ключових особливостей наївного Баєсівського класифікатору є припущення про незалежність ознак. Це означає, що метод вважає всі слова в тексті незалежними один від одного, що на практиці не завжди відповідає реальності. Проте, незважаючи на це спрощення, метод демонструє високу ефективність при вирішенні задач текстової класифікації, зокрема для виявлення спаму, аналізу тональності та автоматизованої модерації коментарів.

Формула Баєса:

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)} \tag{5.1}$$

де $P(c|x)$ – апостеріорна ймовірність належності тексту до класу c за наявності ознак x;

$P(x|c)$ – ймовірність того, що текст з ознаками x зустрічається в класі c;

$P(c)$– апріорна ймовірність класу c (ймовірність зустріти текст цього класу в цілому);

$P(x)$ – нормувальний коефіцієнт, що відображає загальну ймовірність ознак x.

Цей метод називають «наївним», бо припущення про незалежність ознак є спрощенням, яке рідко відповідає реальним текстовим даним. Наприклад, слова у реченні часто взаємопов'язані, і їх значення залежить від контексту. Але ця «наївність» дозволяє значно спростити обчислення і робить метод надзвичайно швидким.

Переваги наївного Баєса:

- висока швидкість навчання і передбачення;
- простота реалізації;
- ефективний навіть на невеликих наборах даних;
- добре працює в умовах дисбалансу класів.

Недоліки:

- ігнорування залежностей між ознаками;
- чутливість до шуму у даних;
- менша ефективність при аналізі тексту з багатозначними словами або саркастичними висловлюваннями.

Наївний Баєс добре працює в умовах, коли потрібно швидко класифікувати великі обсяги тексту, наприклад, при фільтрації спаму або виявленні токсичних коментарів. Метод ефективний для багатокласової класифікації, де існує потреба визначити один із багатьох можливих класів для текстового документа.

4.1.2 Метод опорних векторів (SVM)

Метод опорних векторів є одним із найбільш потужних алгоритмів для задач класифікації тексту, особливо коли дані мають велику кількість ознак. Основна ідея

SVM полягає у побудові гіперплощини, яка максимально розділяє дані різних класів, забезпечуючи максимальну відстань між опорними векторами цих класів.

Цей підхід дозволяє SVM ефективно працювати навіть з високовимірними даними, такими як текстові представлення у вигляді TF-IDF векторів або word embeddings.

Формула для лінійного ядра:

$$f(x) = w^T x + b \tag{5.2}$$

де w – ваговий вектор, який визначає нахил гіперплощини;

x – вхідний вектор ознак;

b – зміщення, що визначає положення гіперплощини у просторі.

Класифікація проводиться за правилом:

$$class(x) = sgn(w^T x + b) \tag{4.3}$$

де w – ваговий вектор, який визначає нахил гіперплощини;

x – вхідний вектор ознак;

b – зміщення, що визначає положення гіперплощини у просторі;

sgn – функція знака числа.

Переваги SVM:

- висока точність на складних даних;
- стійкість до переобучення на великих наборах даних;
- підтримка нелінійних класифікаторів через використання ядер.

Недоліки:

- висока обчислювальна складність при великих обсягах даних;
- чутливість до вибору параметрів та масштабування ознак.

Однією з важливих особливостей SVM є здатність працювати з нелінійно роздільними даними завдяки використанню ядерних функцій (*англ. kernel functions*). Ці функції дозволяють перетворювати простір ознак у більш складний простір вищої розмірності, де дані стають лінійно роздільними.

4.1.3 Логістична регресія

Логістична регресія є одним із найпопулярніших методів для задач бінарної класифікації. Хоча цей метод відносно простий, він демонструє хорошу ефективність у текстовій класифікації та є базовим методом для багатьох NLP-завдань, таких як визначення тональності тексту або виявлення токсичних коментарів.

Логістична регресія передбачає, що залежність між ознаками та ймовірністю належності тексту до певного класу є лінійною. Проте на відміну від лінійної регресії, цей метод застосовує сигмоїдну функцію для перетворення вихідного значення у ймовірність.

Математична основа:

$$\mathrm{P}(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{5.4}$$

де: $\mathrm{P}(y = 1|x)$ – ймовірність належності тексту до класу 1 (токсичний контент);

w – ваговий вектор;

x – вхідний вектор ознак;

b – зміщення.

Переваги:

- простота та швидкість навчання;
- легка інтерпретація вагових коефіцієнтів;
- стійкість до переобучення при використанні регуляризації.

Недоліки:

- ефективна лише для лінійно роздільних даних;
- менш ефективна для нелінійних задач класифікації.

Логістична регресія добре підходить для класифікації лінійно роздільних даних, однак її можна розширити до нелінійних задач за допомогою поліноміальних ознак або інших методів розширення простору ознак.

4.2 Вибір комбінованого методу

В якості мета-моделі для комбінованого ансамблевого методу було обрано градієнтний бустинг. Цей вибір обумовлений кількома факторами:

- здатність вловлювати складні патерни: градієнтний бустинг є методом ансамблевого навчання, який послідовно навчає нові моделі для коригування помилок попередніх;
- стійкість до переобучення: цей метод добре працює з великими наборами даних, що важливо при модерації контенту;
- висока продуктивність: градієнтний бустинг часто демонструє кращі результати порівняно з класичними методами, особливо при складних задачах класифікації тексту.

У наступному розділі будуть представлені результати експериментів та порівняння ефективності різних методів.

4.3. Метрики оцінки методів

Метрики відіграють ключову роль у визначенні ефективності алгоритмів машинного навчання. У процесі модерації текстового контенту критично важливо оцінити, наскільки точно та коректно модель виконує класифікацію. У цьому дослідженні застосовуються такі базові метрики:

4.3.1 Точність (Accuracy)

Точність є однією з найпростіших і найбільш часто використовуваних метрик для оцінки класифікації. Вона показує, яка частка всіх передбачень моделі є правильними.

Формула для розрахунку точності виглядає так:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.5)$$

де: TP (True Positive) – кількість правильно класифікованих токсичних коментарів;

TN (True Negative) – кількість правильно класифікованих нетоксичних коментарів;

FP (False Positive) – кількість хибнопозитивних класифікацій (нетоксичний текст класифікований як токсичний);

FN (False Negative) – кількість хибнонегативних класифікацій (токсичний текст класифікований як нетоксичний).

Точність є хорошою метрикою для збалансованих наборів даних, де кількість позитивних та негативних прикладів є приблизно рівною. Однак при дисбалансі класів ця метрика може бути оманливою, адже висока точність може досягатися шляхом простого ігнорування рідкісного класу.

4.3.2 Влучність (Precision)

Влучність є важливою метрикою для оцінки того, наскільки модель точна у передбаченні позитивного класу. Вона показує, яка частка передбачених токсичних коментарів дійсно є токсичними.

Формула для розрахунку влучності:

$$Precision = \frac{TP}{TP + FP} \quad (5.6)$$

де: TP – кількість коректно класифікованих токсичних коментарів;

FP – кількість хибнопозитивних класифікацій.

Влучність є особливо важливою, коли вартість хибнопозитивної помилки висока. Наприклад, у модерації контенту надмірна кількість хибнопозитивів може призвести до блокування невинних користувачів.

4.3.3 Повнота (Recall)

Повнота оцінює, наскільки добре модель виявляє всі позитивні приклади серед усіх наявних. Іншими словами, вона показує, яка частка токсичних коментарів була успішно ідентифікована.

Формула для обчислення повноти:

$$Recall = \frac{TP}{TP + FN} \tag{5.7}$$

де: TP – кількість правильно класифікованих токсичних коментарів;

FN – кількість хибнонегативних класифікацій (токсичні коментарі, які модель не змогла виявити).

Висока повнота важлива у випадках, коли пропуск токсичного контенту є критичним, наприклад, при модерації вмісту на платформах з високим ризиком небезпечної поведінки.

4.3.4 F1-міра

F1-міра є середнім гармонічним між влучністю та повнотою. Вона використовується для досягнення балансу між двома метриками, коли важливо враховувати як точність, так і здатність моделі виявляти всі позитивні приклади.

Формула для розрахунку F1-міри:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.8}$$

F1-міра є особливо корисною для роботи з незбалансованими наборами даних, де важливо не лише правильно класифікувати більшість прикладів, а й не пропускати рідкісні випадки токсичного контенту.

4.3.5 Методика багатокритеріального порівняння методів

Для визначення найефективнішої моделі буде застосовано метод TOPSIS (*англ. Technique for Order Preference by Similarity to Ideal Solution*), який є одним з найпоширеніших підходів у багатокритеріальному аналізі.

Суть методу полягає в оцінці відносної близькості кожного методу до умовно ідеального рішення, що характеризується максимальними значеннями всіх критеріїв, та антиідеального рішення, тобто з мінімальними значеннями. Метод, який є найближчим до ідеального вектора і далі від найгіршого, вважається кращим.

Першим етапом є нормалізація значень метрик, оскільки вони можуть бути виражені в різних одиницях виміру. Нормалізоване значення критерію обчислюється як:

$$v_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \tag{5.9}$$

де x_{ij} – значення метрики j для методу i;

v_{ij} – нормалізоване значення;

m – кількість методів.

Це забезпечує приведення всіх критеріїв до єдиної шкали, зберігаючи пропорційні відносини між ними.

Після нормалізації обчислюються відстані до ідеального та антиідеального рішень за формулами.

$$D_i^+ = \sqrt{\sum_{j=1}^{n} w_j^2 * \left(v_{ij} - v_j^+\right)^2} \tag{5.10}$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} w_j^2 * \left(v_{ij} - v_j^-\right)^2} \tag{5.11}$$

де v_j^+ – найкраще (максимальне) значення метрики j серед усіх методів;

v_j^- – найгірше (мінімальне) значення метрики j серед усіх методів;

w_j – ваговий коефіцієнт метрики j.

На основі відстаней обчислюється індекс ефективності методу:

$$C_i = \frac{D_i^-}{D_i^{\mp} + D_i^-} \tag{5.12}$$

Для цього експерименту було обрано такі вагові коефіцієнти для метрик:

Таблиця 4.1 – Вагові коефіцієнти метрик (таблиця виконана самостійно)

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

Повноту було враховано з найбільшою вагою (40%), оскільки в задачі модерації контенту найважливіше знайти усі токсичні коментарі.

5 ПРОВЕДЕННЯ ЕКСПЕРИМЕНТУ

У цьому розділі наведено результати проведених експериментів для оцінки продуктивності кожного з алгоритмів класифікації токсичних коментарів. Дослідження включає оцінку базових методів машинного навчання, а саме наївний Баєссіввський класифікатор, метод опорних векторів, логістична регресія, та комбінованого ансамблевого методу на основі градієнтного бустингу.

5.1 Набір даних

Для проведення експерименту використано Toxic Comment Dataset з Kaggle. Цей датасет містить коментарі з позначками токсичності, що дозволяє навчати моделі для автоматизованої модерації текстового контенту.

Ключові характеристики датасету:

- загальна кількість коментарів: ~159 000.
- класифікація: токсичні / нетоксичні коментарі.
- співвідношення класів: значний дисбаланс (переважно нетоксичні коментарі).

Підготовка даних

Серед етапів попередньої обробки тексту важливе місце займає видалення слів, які не несуть змісту, тобто службові частини мови (частки, прийменники, сполучники). Вони не несуть важливої інформації для аналізу тексту і не допомагають під час класифікації. Їх присутність може збільшувати розмірність векторного простору і впливати на ефективність методів машинного навчання.

Наступним етапом є лематизація тексту – процес перетворення слів до їх базової (лематичної) форми. Наприклад, слова «біг», «біжить», «бігали» будуть приведені до однієї форми – «біг». Дана процедура дозволяє зменшити кількість унікальних слів у тексті, що покращує узагальнення і дозволяє алгоритмам ефективніше працювати. Завдяки лематизації модель може краще враховувати

семантичну подібність слів, що позитивно впливає на її здатність виявляти контекст.

Фінальним етапом підготовки даних до навчання є TF-IDF (*англ. Term frequency-inverse document frequency*). Цей процес використовується для перетворення тексту в числове представлення, придатне для обробки алгоритмами машинного навчання. TF-IDF визначає важливість слова відносно всього тексту. Векторизація за допомогою TF-IDF допомагає зменшити вплив загальних слів і підвищити значущість термінів, які краще характеризують текст.

5.3 Результати базових методів

Для навчання та тестування алгоритмів класифікації було написано програму на мові програмування Python, та графічний інтерфейс за допомогою React.

Код скрипту:

```
train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")

X = train['comment_text']
y = train['toxic']

X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

vectorizer = TfidfVectorizer(max_features=5000, stop_words='english')
X_train_vec = vectorizer.fit_transform(X_train)
X_val_vec = vectorizer.transform(X_val)

models = {
    "Naive Bayes": MultinomialNB(),
    "SVM": SVC(kernel='linear', max_iter=1000),
    "Logistic Regression": LogisticRegression(max_iter=1000),
}

for model_name, model in models.items():
    start_time = time.time()

    model.fit(X_train_vec, y_train)

    training_time = time.time() - start_time

    preds = model.predict(X_val_vec)
    print(f"{model_name} результати класифікації:\n", classification_report(y_val, preds))
```

Результати порівняння методів наведені у таблиці 5.1:

Таблиця 5.1 – Результати порівняння методів (таблиця створена самостійно)

Модель	Точність	Влучність	Повнота	F1	Час навчання
Naive Bayes	95%	92%	51%	66%	0.01
SVM	88%	40%	57%	47%	18.63
Logistic Regression	96%	90%	61%	73%	0.23

Аналіз результатів:

- наївний Баєсівський класифікатор показує високу точність (95%), але має низьку повноту (51%), що свідчить про його нездатність виявляти всі токсичні коментарі;
- метод опорних векторів демонструє найгірші результати серед моделей, особливо за влучністю (40%). Однак, повнота (57%) трохи вища, ніж у наївного Баєсівського класифікатора;
- логістична регресія показала найкращий баланс між влучністю (90%) і повнотою (61%), що робить її найбільш продуктивною серед базових моделей.

5.4 Опис комбінованого методу

Для підвищення ефективності методів був створений комбінований ансамблевий класифікатор, який об'єднує три базові методи: наївний Баєсівський класифікатор, метод опорних векторів, логістичну регресію.

У якості мета-моделі використано градієнтний бустинг.

Логіка роботи алгоритму комбінованого методу:

- на першому етапі кожна з трьох базових методів генерує ймовірності для кожного класу;
- ймовірності передаються до мета-моделі, яка приймає остаточне рішення;

– градієнтний бустинг навчається на векторах ймовірностей для покращення класифікації токсичних коментарів.

Реалізовано ансамблевий метод було також за допомогою Python. Код методу:

```
class CustomStackingClassifier(BaseEstimator, ClassifierMixin):
    def __init__(self, alpha=1.0, C=1.0, max_iter=1000, meta_model=None, cv_folds=5, random_state=42):
        self.alpha = alpha
        self.C = C
        self.max_iter = max_iter
        self.cv_folds = cv_folds
        self.random_state = random_state

        self.nb = MultinomialNB(alpha=alpha)
        self.lr = LogisticRegression(max_iter=max_iter, random_state=random_state, solver='liblinear')
        self.svm = SVC(kernel='linear', C=C, probability=True, max_iter=max_iter, random_state=random_state)

        self.meta_model = meta_model if meta_model else GradientBoostingClassifier(
            n_estimators=150,
            learning_rate=0.1,
            max_depth=3,
            random_state=random_state
        )

        self.scaler = StandardScaler()

    def fit(self, X, y):
        print("Навчання базових методів...")
        self.nb.fit(X, y)
        self.lr.fit(X, y)
        self.svm.fit(X, y)

        print("Генерація передбачень через кросс-валідацію...")
        cv = StratifiedKFold(n_splits=self.cv_folds, shuffle=True, random_state=self.random_state)

        nb_preds = cross_val_predict(self.nb, X, y, cv=cv, method='predict_proba')
        lr_preds = cross_val_predict(self.lr, X, y, cv=cv, method='predict_proba')
        svm_preds = cross_val_predict(self.svm, X, y, cv=cv, method='predict_proba')

        stacked_features = np.hstack((nb_preds, lr_preds, svm_preds))
        stacked_features_scaled = self.scaler.fit_transform(stacked_features)

        print("Навчання мета-моделі...")
        self.meta_model.fit(stacked_features_scaled, y)
```

```
            return self

        def predict(self, X):
            nb_preds = self.nb.predict_proba(X)
            lr_preds = self.lr.predict_proba(X)
            svm_preds = self.svm.predict_proba(X)

            stacked_features = np.hstack((nb_preds, lr_preds, svm_preds))
            stacked_features_scaled                                    =
self.scaler.transform(stacked_features)

            return self.meta_model.predict(stacked_features_scaled)

        def predict_proba(self, X):
            nb_preds = self.nb.predict_proba(X)
            lr_preds = self.lr.predict_proba(X)
            svm_preds = self.svm.predict_proba(X)
            stacked_features = np.hstack((nb_preds, lr_preds, svm_preds))
            stacked_features_scaled                                    =
self.scaler.transform(stacked_features)
            return self.meta_model.predict_proba(stacked_features_scaled)
```

5.5 Результати комбінованого методу

Після проведення експерименту було отримано наступні результати:

Таблиця 5.2 – Результати комбінованого методу (таблиця створена самостійно)

Модель	Точність	Влучність	Повнота	F1	Час навчання
Gradient Boosting	96%	84%	78%	76%	241.88

Аналіз результатів:

- комбінована ансамблева модель показує найкращий результат за повнотою (78%) та F1-мірою (76%);
- час навчання значно вищий (241.88 с), однак підвищена продуктивність виправдовує витрати часу.

5.6 Багатокритеріальна оцінка результатів

На основі нормалізованих метрик і вагових коефіцієнтів, було сформовано підсумкову таблицю з результатами моделей класифікації:

Таблиця 5.3 – Багатокритеріальна оцінка результатів (виконано самостійно)

Модель	Точність	Влучність	Повнота	F1	Час навчання	Індекс TOPSIS
Naive Bayes	95%	92%	51%	66%	0.01	0.50
SVM	88%	40%	57%	47%	18.63	0.13
Logistic Regression	96%	90%	61%	73%	0.23	0.64
Gradient Boosting	96%	84%	78%	76%	241.88	0.90

В результаті моделювання комбінований ансамблевий метод на основі градієнтного бустингу отримав найвищу оцінку близькості до ідеального рішення, що підтверджує його ефективність порівняно з базовими методами.

Отже, комбінований ансамблевий метод з використанням градієнтного бустингу показав найкращі результати серед усіх моделей. Незважаючи на довгий час навчання, він досягнув значного покращення повноти та F1-міри, що є критично важливим для задач модерації контенту.

Базові методи, такі як наївний Баєсівський класифікатор та логістична регресія, залишаються ефективними для швидкого розгортання, але їх ефективність значно поступається ансамблевому методу.

5.7 Демонстрація роботи програмного застосунку

Одним із важливих викликів при побудові систем автоматизованої модерації текстового контенту є мовний бар’єр. Більшість існуючих методім машинного

навчання, у тому числі й побудована в межах цього дослідження ансамблева модель, тренуються переважно на англомовних датасетах (наприклад, Toxic Comment Dataset). Через це їх застосування до контенту іншими мовами без додаткової адаптації може бути неефективним.

Для вирішення цього питання у межах розробленого програмного продукту було реалізовано програму для демонстрації того, як можна частково подолати мовний бар'єр за допомогою автоматичного машинного перекладу (рис. 5.1).

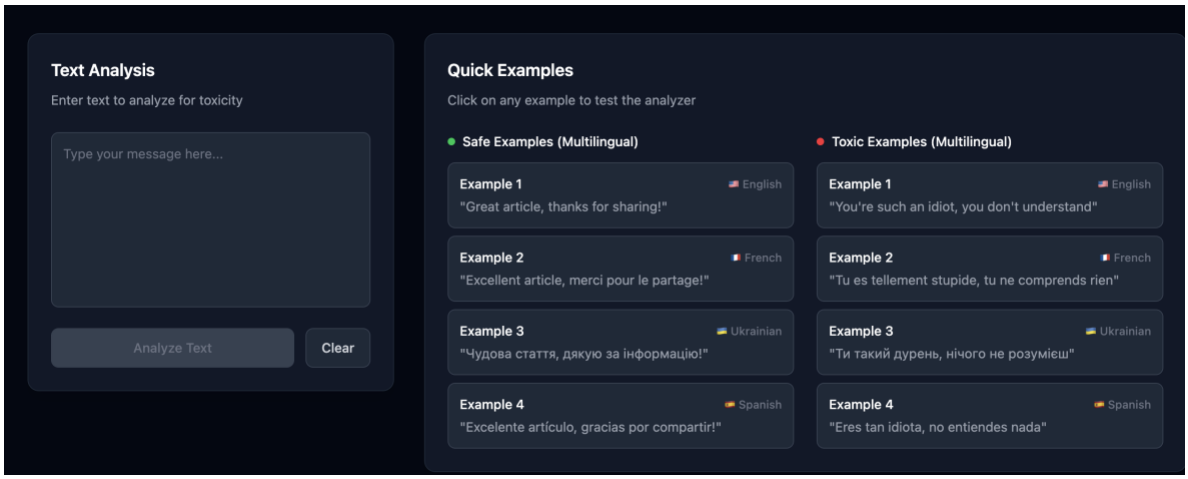

Рисунок 5.1 – Владка «Text Analysis» (виконано самостійно)

На вкладці «Text Analysis» користувач може ввести довільний текст будь-якою мовою. Система автоматично визначає мову введення та виконує його машинний переклад на англійську мову. Уже перекладений текст передається на вхід класифікаторів, побудованих у рамках дослідження.

На рисунках 5.2 та 5.3 наведено приклади роботи системи з текстами різного характеру. На першому прикладі (рис. 5.2) продемонстровано обробку нетоксичного тексту французькою мовою. Після автоматичного перекладу всі моделі — Naive Bayes, SVM, Logistic Regression та Gradient Boosting — класифікували його як безпечний. На другому прикладі (рис. 5.3) показано результат аналізу токсичного тексту українською мовою, який також був правильно розпізнаний як токсичний усіма моделями після перекладу. Слід зазначити, що в обох випадках найточнгший рівень ймовірності токсичності

демонстрував саме градієнтний бустинг, що вказує на його вищу впевненість та стабільність порівняно з іншими базовими методами.

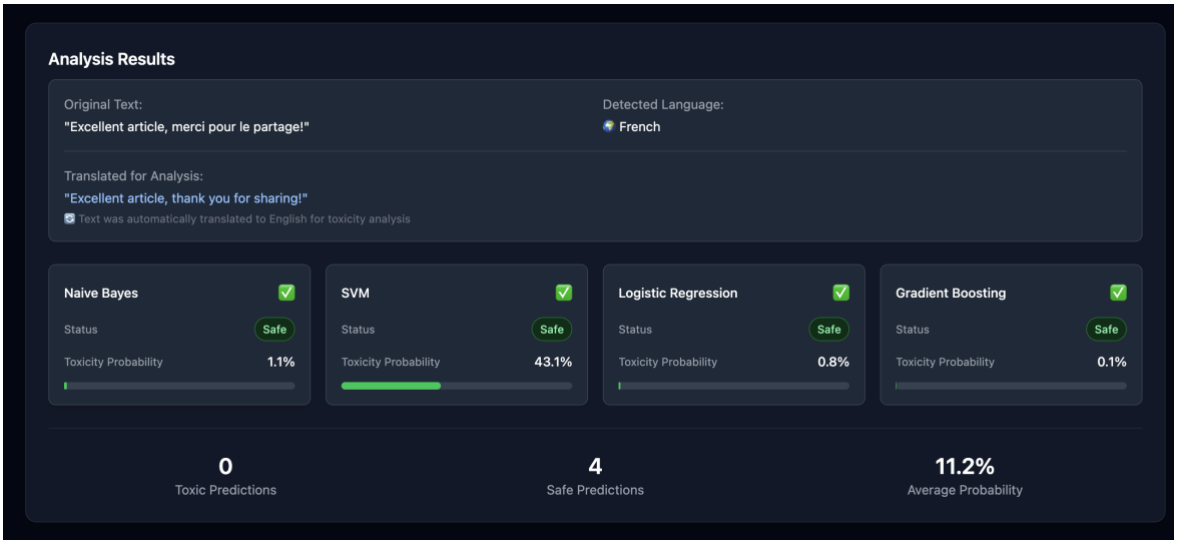

Рисунок 5.2 – Приклад аналізу «нетоксичного» тексту (виконано самостійно)

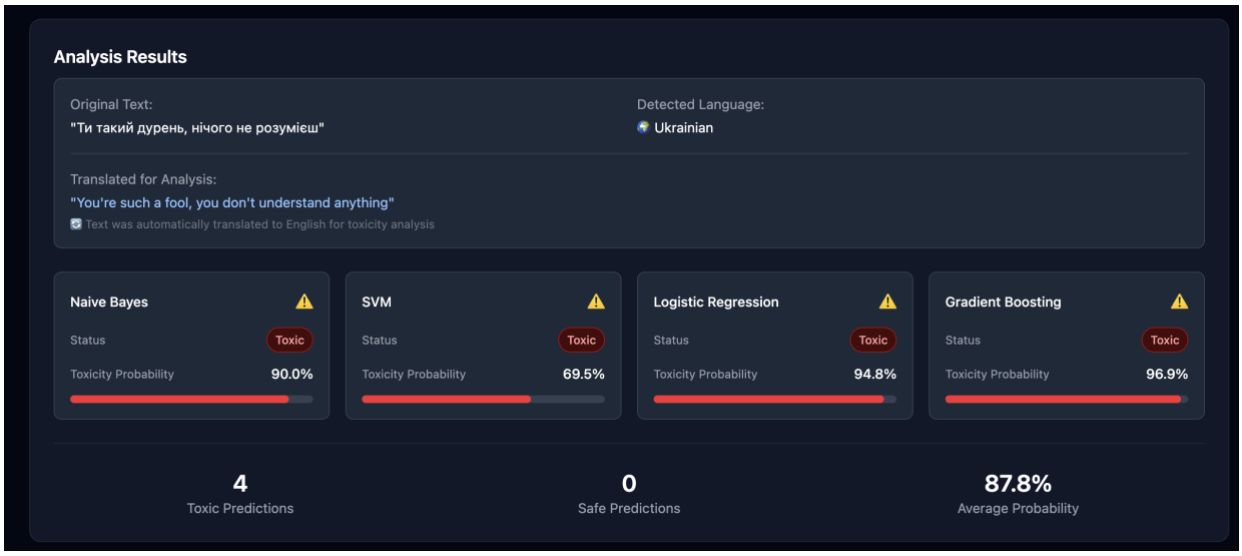

Рисунок 5.3 – Приклад аналізу «токсичного» тексту (виконано самостійно)

Варто зазначити, що такий підхід має певні обмеження: якість класифікації залежить від точності машинного перекладу, а також можливі спотворення контексту. Проте навіть у такому базовому вигляді рішення демонструє

ефективність і дозволяє значно розширити сферу застосування методу для багатомовних платформ.

У перспективі, подальший розвиток цього напрямку може включати використання мультимовних моделей або тонке донавчання на текстах інших мов, що дозволить підвищити якість модерації без залежності від проміжного перекладу.

ВИСНОВКИ

У ході виконання роботи були досягнуті всі поставлені задачі. Було проведено детальний аналіз існуючих методів автоматизованої модерації текстового контенту, включаючи класичні методи, такі як наївний Баєсівський класифікатор, метод опорних векторів та логістична регресія. Аналіз показав, що класичні методи демонструють високу продуктивність на простих наборах даних, але можуть бути покращені, шляхом використання їх у ансамблевих методах для навчання мета-моделі. Частину результатів цього дослідження було представлено у вигляді статті на Всеукраїнській науково-практичній конференції студентів та молодих вчених "Радіоелектроніка та молодь у XXI столітті" (ХНУРЕ, 2025) [16].

У ході тестування базових методів було виявлено, що наївний Баєсівський класифікатор показав високу точність при швидкому навчанні, однак його низький рівень повноти вказує на значну кількість пропущених токсичних коментарів. Метод опорних векторів продемонстрував помірну продуктивність, але мав найнижчу точність та влучність. Логістична регресія виявилась найбільш збалансованою серед базових моделей і продемонструвала найкраще співвідношення між влучністю та повнотою.

У рамках дослідження було розроблено та реалізовано комбінований ансамблевий метод, який поєднує результати роботи базових методів із використанням градієнтного бустингу як мета-моделі. Це дозволило підвищити ефективність класифікації за рахунок компенсації недоліків окремих алгоритмів. Також було розроблено графічний інтерфейс для демонстрації результатів експерименту.

Крім того, у межах розробленого програмного продукту було реалізовано додатковий функціонал для часткового подолання мовного бар’єру шляхом автоматичного машинного перекладу вхідного тексту перед передачею до класифікаторів. Цей підхід дозволив продемонструвати можливість ефективного застосування методу до багатомовного контенту без потреби повного донавчання на кожній окремій мові. Подальший розвиток цієї ідеї може включати інтеграцію

мультимовних моделей та оптимізацію процесу перекладу для підвищення якості модерації. Код усіх артефактів, отриманих у результаті розробки завантажено на Github [17].

Запропонований метод було протестовано на наборі даних Toxic Comment Dataset з платформи Kaggle. 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.

Однак під час дослідження були виявлені й певні обмеження. Значні обчислювальні витрати для навчання ансамблевого методу стали одним із викликів у процесі роботи. Навіть на достатньо урізаному наборі тестових даних, ансамблева модель перевищила швидкість навчання у десятки разів у порівнянні з базовими моделями. Потенційна упередженість моделі, пов’язана з дисбалансом даних у наборі Toxic Comment Dataset, залишається ще однією проблемою, яку необхідно вирішити для підвищення ефективності модерації.

Подальші напрями роботи можуть включати інтеграцію трансформерних моделей (BERT, DistilBERT) як мета-моделей для ансамблю, дослідження мультимодальних підходів для модерації тексту та візуального контенту одночасно, а також розробку методів зменшення упередженості моделі та покращення її адаптації до нових доменів.

Загалом, результати роботи підтверджують ефективність ансамблевого методу та демонструють доцільність його застосування для автоматизованої модерації текстового контенту в цифрових середовищах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Building ML models for everyone: understanding fairness in machine learning [Електронний ресурс] // Google AI Blog. – 2019. – URL: https://cloud.google.com/blog/products/ai-machine-learning/building-ml-models-for-everyone-understanding-fairness-in-machine-learning (дата звернення: 27.02.2025).

2. Self-supervised Learning for Language Understanding [Електронний ресурс] // arXiv. – 2020. – URL: https://arxiv.org/abs/2005.12766 (дата звернення: 10.03.2025).

3. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Електронний ресурс] // arXiv. – 2018. – URL: https://arxiv.org/abs/1810.04805 (дата звернення: 10.03.2025).

4. Jigsaw Toxic Comment Classification Challenge [Електронний ресурс] // Kaggle. – 2018. – URL: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (дата звернення: 10.03.2025).

5. Text Classifiers in Machine Learning [Електронний ресурс] // IEEE Xplore. – 2024. – URL: https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guide (дата звернення: 11.03.2025).

6. Jigsaw Unintended Bias in Toxicity Classification [Електронний ресурс] // Kaggle. – 2019. – URL: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification (дата звернення: 11.03.2025).

7. Reducing Bias and Improving Fairness in Machine Learning Models [Електронний ресурс] // deepchecks. – 2023. – URL: https://www.deepchecks.com/reducing-bias-and-ensuring-fairness-in-machine-learning (дата звернення: 20.03.2025).

8. Sharonova, N., Kyrychenko, I., Gruzdo, I., Tereshchenko, G. Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types CEUR Workshop Proceedings, 2022, 3171, pp. 16–26.

9. The State of Online Harassment [Електронний ресурс] // Pew Research Center. – 2021. – URL: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (дата звернення: 20.03.2025).

10. Global Online Safety Survey [Електронний ресурс] // Microsoft. – 2024. – URL: https://www.microsoft.com/en-us/digitalsafety/research/global-online-safety-survey (дата звернення: 22.03.2025).

11. Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models [Електронний ресурс] // ACL Anthology. – 2024. – URL: https://aclanthology.org/2024.findings-emnlp.254/ (дата звернення: 22.03.2025).

12. Використання алгоритмів машинного навчання для автоматизації процесу модерації контенту в групових чатах месенджерів [Електронний ресурс] / Мокрицька О. В., Мочернюк Ю. М. – 2024. URL: https://nv.nltu.edu.ua/index.php/journal/article/view/2669 (дата звернення: 22.03.2025).

13. Ganaie M. A., Hu M., Tanveer M., Suganthan P. N. Ensemble deep learning: A review [Електронний ресурс] // arXiv. – 2022. – URL: https://arxiv.org/abs/2104.02395 (дата звернення: 22.03.2025).

14. Kowsari K., Meimandi K. J., Heidarysafa M., Mendu S., Barnes L. E., Brown D. Text Classification Algorithms: A Survey [Електронний ресурс] // arXiv. – 2019. – URL: https://arxiv.org/abs/1904.08067 (дата звернення: 24.03.2025).

15. Безродний В.В., Турута О.П. Дослідження методів виявлення синтетичних текстів, 27-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. матеріалів форуму. Т. 6., Ч. І. – Харків: ХНУРЕ. 2023. – 420 с.

16. Керецман І.А. Модерація текстового контенту з використанням комбінованого ансамблевого підходу на основі алгоритмів машинного навчання, 29-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». (Харків, 16-19 квітня 2025 року). Зб. Матеріалів форуму. Т.6., - Харків: ХНУРЕ. 2025. – 630 с.

17. Github – 2025_M_ПІ_ІПЗ-23-2_Керецман_І_А URL: https://github.com/illiakeretsman/2025_M_PI_IPZ-23-2_Keretsman_I_A (дата звернення: 11.06.2025)

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

8. Sharonova, N., Kyrychenko, I., Gruzdo, I., Tereshchenko, G. Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types CEUR Workshop Proceedings, 2022, 3171, pp. 16–26.

16. Керецман І.А. Модерація текстового контенту з використанням комбінованого ансамблевого підходу на основі алгоритмів машинного навчання 29-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». (Харків, 16-19 квітня 2025 року). Зб. Матеріалів форуму. Т.6., - Харків: ХНУРЕ. 2025. – 630 с.