ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

УДК 004.9

ОБРОБКА ТЕКСТОВИХ ДАНИХ СОЦІАЛЬНИХ МЕДІА НА ПРИРОДНІЙ МОВІ ЗА ДОПОМОГОЮ BERT ТА XGBOOST

Батюк Т. М. – аспірант кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

Досин Д. Г. – д-р техн. наук, старший науковий співробітник, професор кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Зростання обсягу текстових даних у соціальних мережах вимагає розробки ефективних методів аналізу настроїв, здатних враховувати як лексичні, так і контекстуальні залежності. Традиційні підходи до обробки тексту мають обмеження у розумінні семантичних зв'язків між словами, що впливає на точність класифікації. Інтеграція глибоких нейронних мереж для векторизації тексту з ансамблевими алгоритмами машинного навчання та методами інтерпретації результатів дозволяє покращити якість аналізу настроїв.

Метою дослідження є розробка та оцінка нового підходу до класифікації настроїв текстових повідомлень, що поєднує Sentence-BERT для глибокої семантичної векторизації, XGBoost для високоточної класифікації, SHAP для пояснення внеску ознак, sentence embedding similarity для оцінки семантичної подібності та λ-регуляризацію для покращення узагальнюючої здатності моделі. Дослідження спрямоване на аналіз впливу цих методів на якість класифікації, визначення найбільш значущих ознак та оптимізацію параметрів для забезпечення балансу між точністю та інтерпретованістю моделі.

Метод. У дослідженні використовується Sentence-BERT для перетворення текстових даних у векторний простір із глибокими семантичними зв'язками. Для класифікації настроїв застосовується XGBoost, який забезпечує високу точність та стабільність навіть на нерівномірно розподілених наборах даних. Для пояснення внеску ознак використано метод SHAP, що дозволяє визначити, які фактори найбільше впливають на прогноз. Додатково використовується sentence embedding similarity для порівняння текстів за семантичною подібністю, а λ-регуляризація оптимізує баланс між узагальненням та точністю моделі.

Результати. Запропонований підхід демонструє високу ефективність у задачах класифікації настроїв. Значення ROC-AUC підтверджує здатність моделі точно розрізняти класи емоційного забарвлення тексту. Використання SHAP забезпечує інтерпретованість результатів, дозволяючи пояснити вплив кожної ознаки на класифікацію. Sentence embedding similarity підтверджує ефективність Sentence-BERT у виявленні семантично подібних текстів, а λ-регуляризація покращує узагальнюючу здатність моделі.

Висновки. Дослідження демонструє наукову новизну через комплексне поєднання Sentence-BERT, XGBoost, SHAP, sentence embedding similarity та λ-регуляризації для покращення точності та інтерпретованості аналізу настроїв. Отримані результати підтверджують ефективність запропонованого підходу, що робить його перспективним для застосування у моніторингу громадської думки, автоматизованій модерації контенту та персоналізованих рекомендаційних системах. Подальші дослідження можуть бути спрямовані на адаптацію моделі до специфічних доменів, розширення джерел текстових даних та вдосконалення методів інтерпретації для покращення довіри до автоматизованого аналізу настроїв.

КЛЮЧОВІ СЛОВА: Машинне навчання, нормалізація ознак, трансформери, матриця плутанини, Sentence-BERT, класифікація текстових даних.

АБРЕВІАТУРИ

SB – sentence-BERT;
CB – XGBoost;
ВП – векторний простір;
ВТ – векторизація тексту;
НФ – нормалізація функції;
ПО – поєднання ознак;
ПГ – посилення градієнта;
ОМ – оцінка моделі;
SNS – Sentence Embedding Similarity.

НОМЕНКЛАТУРА

$\cos(\theta)$ – косинусна схожість;
A – вектор першого значення;
B – вектор другого значення;
A_i – компоненти вектору A;
B_i – компоненти вектору B;
$|D|$ – кількість даних у поточній колекції;
f_{BERT} – функція трансформера S-BERT;
v – вектор фіксованої довжини;
T – текстовий вхід;
ε – фактор нульового вектора;
λ – коефіцієнт семантичної подібності;
μ – середнє значення ознаки;
σ – стандартне відхилення ознаки;
n_i – загальна кількість повторень слова;

DOI 10.15588/1607-3274-2025-2-14

$\sum_k n_k$ – загальна кількість слів у документі;
$\sum A_j B_j$ – скалярний добуток векторів A та B;
$\sum \lambda |A_j||B_j|$ – компонент, що враховує вплив модулів елементів;
$\sum (A_j^2)^{1\backslash 2}$ – нормалізація вектора A;
$\sum (B_j^2)^{1\backslash 2}$ – нормалізація вектора B;
n_j – середня кількість повторень слова;
F_μ – підхід, який описує оцінку ознак;
F_σ – підхід, що описує стандартне відхилення оцінки ознаки;
n – розмірність векторів A і B;
μ_1 – вектори, що представляють два документи;
μ_2 – скалярний добуток векторів;
S_i – об'єкт описаний векторним набором значень;
i – індекс компоненту модулів елементів вектора;
j – індекс векторного елементу;
N – множина повторень елементів вектора;
M – множина параметрів моделі даних;
Σ – множина стандартних відхилень ознак;
Θ – множина коефіцієнтів векторних перетворень;
X – система аналізу текстових даних;
n_k – кількість слів в розподіленому документі;
k – індекс векторного розподілу;

ВСТУП

У сучасну епоху цифровізації стрімке зростання обсягу текстових даних у соціальних мережах створює нові виклики для їх обробки та аналізу. Соціальні платформи, такі як Twitter, є важливим джерелом інформації для вивчення громадської думки, прогнозування поведінкових трендів і моніторингу соціальних процесів. Однак, аналіз цих даних ускладнюється їх неструктурованим характером, варіативністю мовних конструкцій та високою швидкістю оновлення контенту. Традиційні методи обробки тексту часто виявляються недостатньо ефективними для розв'язання цих задач, що обумовлює необхідність розробки нових підходів.

Актуальність дослідження зумовлена потребою у вдосконалених методах аналізу текстових даних, які здатні точно визначати емоційний тон повідомлень та забезпечувати їх ефективну класифікацію. Сучасні трансформерні моделі, зокрема Sentence-BERT, дозволяють отримати якісні векторні представлення тексту, що враховують як лексичний, так і контекстуальний рівні аналізу. Однак, застосування стандартного Sentence-BERT до аналізу настроїв у соціальних мережах потребує адаптації та вдосконалення його механізмів, оскільки класичні підходи не враховують специфічні особливості коротких текстових повідомлень, таких як лексичні скорочення, емодзі та багатозначність виразів.

Предметом дослідження є методи аналізу текстових даних соціальних мереж, а об'єктом – процес векторизації тексту та його подальша класифікація за допомогою алгоритмів машинного навчання. У межах дослідження розглядаються сучасні методи обробки природної мови, що включають застосування глибоких нейронних мереж для створення числових представлень тексту та використання алгоритмів градієнтного бустингу для їх аналізу. Особливу увагу приділено удосконаленню методів векторизації тексту, які дають змогу отримати більш точне представлення семантичного змісту повідомлень.

Для цього Sentence-BERT проходить додаткове навчання на текстових даних соціальних мереж, що дозволяє враховувати їх специфіку та покращує якість отриманих векторних ознак. Розробка нової стратегії агрегованого пулінгу дозволяє підвищити точність обчислення семантичної близькості між текстами, оскільки враховує не лише середнє значення вихідних представлень слів, а й локальні особливості їх використання в контексті.

Додатково здійснюється адаптація алгоритму XGBoost до роботи з текстовими ознаками, що включає розширення набору вхідних характеристик, а також використання методів балансування класів для підвищення стійкості моделі до нерівномірного розподілу даних у навчальному наборі. Завдяки комплексному підходу, що включає поєднання трансформерних моделей та градієнтного бустингу, запропоноване рішення сприяє підвищенню якості класифікації настроїв у текстах соціальних мереж.

Основною метою роботи є розробка вдосконаленого підходу до класифікації настроїв текстових повідомлень шляхом модифікації Sentence-BERT та адаптації XGBoost для обробки векторизованих текстових даних. Для досягнення цієї мети необхідно вирішити кілька ключових задач. По-перше, необхідно здійснити додаткове навчання Sentence-BERT на текстах соціальних мереж, що дозволить покращити його здатність враховувати контекстуальні особливості коротких повідомлень. По-друге, слід розробити ефективну стратегію агрегованого пулінгу, яка забезпечить більш точне представлення текстових ознак та сприятиме кращому захопленню семантичних залежностей між словами. По-третє, потрібно адаптувати алгоритм XGBoost для роботи із текстовими ознаками, що передбачає розширення його можливостей щодо обробки категоріальних даних та підвищення його стійкості до можливих аномалій у текстових представленнях. Нарешті, необхідно впровадити методи балансування класів, які допоможуть усунути проблему нерівномірного розподілу даних.

Наукова новизна дослідження полягає у розробці адаптивного підходу до аналізу текстових даних, що поєднує вдосконалену архітектуру Sentence-BERT із глибоким навчанням на текстах соціальних мереж та адаптований механізм обробки ознак у XGBoost. Додавання механізму контрастного навчання дозволяє підвищити чутливість моделі до контекстних особливостей повідомлень. Крім того, адаптація XGBoost до обробки текстових ознак шляхом оптимізації вхідних характеристик та введення додаткових регуляційних стратегій забезпечує стабільну роботу моделі.

DOI 10.15588/1607-3274-2025-2-14

Запропоновані вдосконалення дозволяють підвищити точність класифікації настроїв у коротких текстах, мінімізувати вплив дисбалансу класів та забезпечити стійкість моделі до шумових факторів у даних. Практична суть дослідження полягає в потенційному застосуванні в системах автоматизованого моніторингу громадської думки та аналізу клієнтських відгуків що є важливим аспектом для розвитку сучасних інформаційно-аналітичних систем.

Внесок роботи полягає у розробці нових підходів до вдосконалення трансформерних моделей у сфері аналізу текстових даних соціальних мереж та оптимізації алгоритмів градієнтного бустингу для задач текстової класифікації.

1 ПОСТАНОВКА ПРОБЛЕМИ

Інформаційна система аналізу текстових даних з використанням алгоритмів векторизації Sentence-BERT представлена імітаційною моделлю кортежу:

$$X = <\cos(\theta),\ A,\ B,\ |D|,\ \sum_k n_k,\ N,\ F\mu,\ F\sigma >,$$

де $N = \{n_1, n_2, n_3, n_4\}$, $M = \{\mu_1, \mu_2, \mu_3, \mu_4\}$, $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$, $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$.

XGBoost – це потужна техніка машинного навчання, яка використовується для задач регресії та класифікації. Його основна ідея полягає в побудові моделі шляхом поступового додавання нових моделей, які виправляють помилки попередніх:

$$\cos(\theta) = A_i \circ B_i \circ n_i,$$
$$|D| = n_i(n_j(n(\mu_1, \mu_2, \mu_3), \sigma_1, \sigma_2), \mu_4),$$
$$n = n_i \circ n_j \circ n_k.$$

Кількість дійсних текстових даних у остаточній векторній колекції: $S_i = i \circ j \circ k$, тому

$$A_iB_i = \theta_1(n_1(\mu_1(S_i, n_4, \sigma_4, \mu_4), \theta_2, \mu_3), \theta_3).$$

Загальна кількість слів у документі та зовнішні нормалізовані дані:

$$\sum_k n_k = \mu \circ \sigma \circ \lambda \circ \Sigma,$$
$$N = n_3(n_4(\mu_4(\theta_4(\mu_2, \theta_2), n_1), \mu_1), \mu_3).$$

Підходи, які описують поточну оцінку функції та стандартне відхилення оцінки функції:

$$F\mu = N \circ M \circ \Sigma \circ \Theta,$$
$$F\sigma = \theta_4(\mu_4(\theta_3(\mu_3(A_iB_i, \sum_k n_k), |D|), n_4), n_3).$$

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

У статті [1] авторами досліджується ефективність застосування моделей глибокого навчання, таких як BERT, для векторизації текстів у задачах класифікації настроїв. Основна увага приділяється порівнянню BERT із класичними підходами, зокрема TF-IDF і Word2Vec. Автори підкреслюють переваги BERT у врахуванні контекстуальних зв'язків між словами, що дозволяє значно покращити точність класифікації. Крім того, дослідження включає оцінку впливу різних параметрів векторизації на результати моделювання, таких як розмір словникового запасу та кількість шарів трансформера.

У роботі [2] акцентується увага на інтеграції XGBoost як основної моделі класифікації в задачах аналізу настроїв. Автори зосереджуються на оптимізації гіперпараметрів моделі та її здатності працювати з векторними представленнями текстів, створеними за допомогою BERT. У статті також розглядається вплив дисбалансу класів на результати моделювання, а для його компенсації застосовуються методи збалансування даних, такі як oversampling і undersampling. Отримані результати демонструють високу точність моделі, особливо у випадках, коли використовуються додаткові категоріальні ознаки, такі як час публікації повідомлення або місце користувачів.

Стаття [3] представляє новаторський підхід до оцінки схожості текстів на основі Sentence Embedding Similarity. Автори використовують косинусну схожість між векторами, створеними Sentence-BERT, для аналізу семантичної близькості між повідомленнями. Це дозволяє виявляти схожі за змістом повідомлення та групувати їх у кластеризовані теми. Основний внесок роботи полягає в демонстрації, як подібність між текстами може бути використана для покращення класифікації настроїв, зокрема, для вирішення неоднозначних випадків, де текст містить змішані емоції.

Дослідження [4] зосереджене на вивченні впливу специфічної лексики на результати класифікації настроїв. Автори досліджують, як окремі ключові слова або фрази, такі як «щасливий», «злий» чи «розчарований», впливають на ймовірність передбачень моделі. Використовуючи аналіз важливості ознак у моделі XGBoost, вони показують, що специфічна лексика має суттєвий вплив на результати класифікації. Робота також включає експерименти з вилученням високочастотних слів і аналізом того, як це змінює точність моделі.

У статті [5] розглядається важливість візуалізації результатів класифікації настроїв. Автори пропонують використовувати різні типи графіків, такі як ROC-криві, матриці плутанини, Precision-Recall криві та теплові карти косинусної схожості, для оцінки продуктивності моделі та інтерпретації результатів. Особливу увагу приділено інтерактивним інструментам візуалізації, які дозволяють аналізувати зв'язки між семантично подібними текстами та їх вплив на результати класифікації.

Дослідження [6] підкреслює важливість комбінування моделей, таких як BERT і XGBoost, у задачах аналізу настроїв. У статті представлено експерименти, де ці моделі використовуються у зв'язці для отримання точніших результатів. Автори доводять, що використання гібридних підходів дозволяє краще враховувати як лексичні, так і контекстуальні особливості текстів, що позитивно впливає на загальну продуктивність наведених в статті моделей.

DOI 10.15588/1607-3274-2025-2-14

Стаття [7] демонструє використання Sentence Embedding Similarity для створення кластерів текстів у реальних наборах даних із соціальних мереж. Автори доводять, що цей підхід може бути корисним для зменшення розмірності даних і виділення ключових тем, які потім використовуються для уточнення передбачень настроїв.

У роботі [8] проводиться порівняльний аналіз різних типів векторних представлень, таких як Sentence-BERT, FastText та GloVe, у контексті класифікації настроїв. Автори роблять висновок, що Sentence-BERT забезпечує найвищу точність, особливо у поєднанні з алгоритмами градієнтного бустингу, такими як XGBoost.

У статті [9] автори пропонують застосування моделі XGBoost для автоматичної класифікації настроїв у соціальних мережах, таких як Twitter.

Основна увага приділяється аналізу факторів, які впливають на продуктивність моделі, зокрема, впливу різних підходів до векторизації текстів. Автори використовують Sentence-BERT для створення текстових векторів, а також досліджують, як різні розміри векторів та методи нормалізації впливають на точність передбачень. Робота підкреслює, що інтеграція векторів із додатковими ознаками, такими як час публікації чи кількість лайків, суттєво підвищує точність класифікації.

У статті [10] акцентується увага на застосуванні Sentence Embedding Similarity для групування текстів за темами. Автори досліджують, як використання косинусної схожості між векторами повідомлень може допомогти в аналізі настроїв у великих наборах даних. Основний внесок цієї роботи полягає у демонстрації ефективності кластеризації повідомлень за допомогою схожості їх текстових представлень. Наприклад, у статті представлено результати експериментів, де повідомлення, які містять однакові ключові слова або фрази, автоматично групуються в кластери для подальшого аналізу. Це дозволяє ефективно зменшувати розмірність даних і зосереджувати увагу на ключових темах, які обговорюються у соціальних мережах.

У статті [11] розглядається питання візуалізації результатів аналізу настроїв за допомогою інтерактивних графічних інструментів. Автори пропонують використовувати теплові карти, які показують косинусну схожість між текстовими векторами, створеними Sentence-BERT, для кращого розуміння зв'язків між повідомленнями. Також у роботі представлено інші типи візуалізацій, зокрема ROC-криві, Precision-Recall криві та бар-чарти, що дозволяють оцінювати точність і повноту моделей класифікації.

Основна увага приділяється інтерактивності таких графіків, які дозволяють дослідникам глибше аналізувати зв'язки між даними, а також виявляти патерни у поведінці користувачів соціальних мереж. Крім того, у статті наводяться приклади використання таких інструментів для ідентифікації фейкових новин або виявлення токсичних коментарів. Результат аналізу наявних наукових робіт демонструє значний прогрес у використанні сучасних методів обробки тексту, таких як BERT, XGBoost і Sentence Embedding Similarity, для класифікації настроїв у соціальних мережах. Ці дослідження показують важливість інтеграції трансформерних моделей для врахування контексту текстів, ансамблевих методів для підвищення точності передбачень та візуалізацій для кращого розуміння результатів. Особливо цінними є підходи до вирішення проблеми дисбалансу класів, аналізу семантичної схожості текстів, впливу специфічної лексики та кластеризації даних. Водночас залишаються відкритими питання щодо подальшого вдосконалення цих методів. Зокрема, потребує уваги оптимізація процесу навчання моделей на великих наборах даних, врахування мовних і культурних відмінностей у текстах.

Крім того, важливим напрямом є дослідження методів пояснення результатів моделей для кращого розуміння, чому модель приймає ті чи інші рішення, що може значно підвищити довіру до автоматизованих систем аналізу настроїв.

3 МАТЕРІАЛИ ТА МЕТОДИ

Це дослідження спрямоване на аналіз настроїв у текстових повідомленнях із соціальних мереж, використовуючи передові методи обробки тексту та глибинного навчання. Для аналізу використано датасет Kaggle Dataset: Sentiment140, що містить повідомлення, помічені як негативні, нейтральні або позитивні, що дозволяє ефективно класифікувати настрій повідомлень за допомогою машинного навчання. Основним завданням є створення моделі, здатної не тільки класифікувати настрій, але й оцінювати семантичну схожість між текстами. Для досягнення цієї мети були використані методи Sentence-BERT та Sentence Embedding Similarity [12]. Було запропоновано удосконалений підхід до оцінки семантичної схожості між текстами, застосовуючи нові методи агрегації векторів, що дозволяють більш точно відображати зміст текстів у порівнянні з класичними методами.

У дослідженні ми використовуємо метод косинусної схожості [13] для вимірювання подібності між текстами, де кожен текст спочатку перетворюється у вектор фіксованої довжини за допомогою моделі Sentence-BERT. Це дозволяє забезпечити точну оцінку схожості текстів не тільки за допомогою наявних слів, але й через семантичні зв'язки та контекстуальну подібність. Моделі трансформерів, такі як BERT [14], використовуються для отримання числових векторних уявлень, що зберігають контекстуальні властивості слів. Після перетворення тексту в таку форму, для оцінки схожості використовується вдосконалена версія Sentence Embedding Similarity. Удосконалена формула для обчислення Sentence Embedding Similarity заснована на врахуванні не тільки косинусної

DOI 10.15588/1607-3274-2025-2-14

схожості між векторами, але й додаткових факторів, таких як контекстуальна інформація та семантична вагомість [15] окремих компонентів векторів. Це дозволяє отримати більш точні результати для порівняння текстів, зокрема в задачах, пов'язаних із класифікацією настроїв та виявленням схожості між повідомленнями у соціальних мережах.

Удосконалена формула для обчислення схожості між двома реченнями, наведена в формулі 1, де A і B – це вектори, що представляють два речення S_1 та S_2 , отримані через модель, таку як Sentence-BERT. A_i та B_i це компоненти векторів A і B, що відображають значення на відповідних позиціях, n – кількість елементів у векторі (розмірність векторів), λ – коефіцієнт, який визначає важливість різниці між величинами компонентів векторів для покращення семантичної подібності, ε – маленьке значення для уникнення ділення на нуль у випадку нульового вектора

$$ses(S_1,S_2)=\frac{\sum_{i=1}^{n}A_jB_j+\lambda\times\sum_{i=1}^{n}|A_j|\times|B_j|}{\sqrt{\sum_{i=1}^{n}A_j^2+\varepsilon}\times\sqrt{\sum_{i=1}^{n}B_j^2+\varepsilon}}. \quad (1)$$

Моделі на основі трансформерів, зокрема Sentence-BERT, набули значної популярності завдяки своїй здатності ефективно перетворювати текст у числовий формат, що враховує семантичний контекст слів у реченнях. Застосування таких моделей відкриває нові можливості для використання текстових даних в алгоритмах машинного навчання, зокрема для класифікації, кластеризації та інших задач аналізу тексту. Sentence-BERT [16] є модифікацією BERT, оптимізованою для обчислення семантичної схожості між текстами, що робить його надзвичайно ефективним для вирішення задач парного порівняння текстів та визначення схожості між ними. Основною метою S-BERT, що наведений в формулі 2, де T – текстовий вхід, f_{BERT} – функція трансформера S-BERT та v – результат у вигляді вектора фіксованої довжини, що відображає зміст текстів. S-BERT використовує додатковий шар агрегації (наприклад, середнє або максимальне значення) для отримання одного вектора на рівні всього тексту. Це дозволяє швидко та ефективно обчислювати семантичну схожість між текстами.

$$v = f_{BERT}(T). \quad (2)$$

Однією з основних характеристик S-BERT є його здатність працювати із завданнями парного порівняння текстів, що важливо для задач, де потрібно оцінити схожість або відмінності між двома текстами. Завдяки можливості використовувати метрики, такі як косинусна схожість, S-BERT може оцінювати не тільки наявність однакових слів, а й їх контекстуальні зв'язки в реченні. Це дозволяє моделі ефективно застосовуватися для завдань, таких як класифікація настроїв або пошук схожих повідомлень в текстових даних. Крім того, S-BERT є надзвичайно корисним для задач кластеризації, оскільки створює векторні репрезентації, що зберігають семантичні зв'язки між різними елементами тексту.

Для того, щоб максимізувати ефективність роботи з текстами, S-BERT поєднує передтреновані моделі трансформерів, такі як BERT чи DistilBERT, з операцією агрегації, що дозволяє скоротити розмір вихідних векторів, не втрачаючи важливих семантичних особливостей тексту.

Це створює потужний інструмент для аналізу тексту, який забезпечує високу точність при розв'язанні задач класифікації настроїв або пошуку подібних текстів. У той же час, обраний підхід дозволяє зберегти точність обробки текстів при зниженні обсягу необхідних для обробки даних, що робить S-BERT практичним та ефективним у багатьох реальних застосунках.

Удосконалення моделі S-BERT полягає в оптимізації її здатності до обробки великих обсягів текстових даних, зокрема завдяки адаптації агрегаційних функцій для отримання більш точних та репрезентативних векторних характеристик.

У результаті, застосування цієї моделі дозволяє не тільки значно покращити точність класифікації або кластеризації, а й суттєво знизити вимоги до обчислювальних ресурсів при виконанні великих обсягів аналізу текстових даних.

Таким чином, запропонований підхід із використанням моделі S-BERT вносить значний внесок, удосконалюючи методи семантичного аналізу текстів, зокрема через покращення якості та ефективності побудови векторних представлень текстових даних. Важливими аспектами цієї роботи є як теоретичні, так і практичні новації, що сприяють підвищенню точності та зниженню витрат ресурсів при обробці великих обсягів тексту в реальних умовах.. Блок-схему роботи моделі S-BERT зображено на рисунку 1.

Рисунок 1 – Блок-схема роботи моделі S-BERT

DOI 10.15588/1607-3274-2025-2-14

Отримавши векторизований текст з семантичним контекстом наступним кроком буде здійснення машинного навчання з використанням XGBoost, потужної бібліотеки для машинного навчання, яка використовується для вирішення завдань класифікації та регресії. Основною перевагою XGBoost є її здатність ефективно працювати з категоріальними ознаками без необхідності попереднього перетворення або кодування, що значно спрощує підготовку даних і прискорює процес навчання.

Крім того, вона використовує передову технологію градієнтного бустінгу, яка забезпечує високу точність і стабільність моделі. XGBoost базується на методі градієнтного бустінгу на деревах рішень (Gradient Boosting on Decision Trees), який є ефективним і популярним підходом для побудови моделей. Цей метод передбачає побудову послідовності дерев рішень, де кожне наступне дерево коригує помилки попереднього.

Категоріальні ознаки в XGBoost обробляються без необхідності їх явного перетворення в числові значення, що зменшує ймовірність втрати інформації і покращує загальну ефективність. XGBoost є особливо ефективним для задач, де потрібно здійснити класифікацію текстів, в тому числі в задачах аналізу настроїв, спаму, класифікації тематики текстів тощо.

Враховуючи можливість інтеграції з іншими техніками обробки природної мови, такими як Sentence-BERT, XGBoost може отримувати високоточні векторні представлення текстів і ефективно використовувати їх для класифікації.

За допомогою моделей, таких як Sentence-BERT, текстові дані перетворюються у високоякісні вектори, що зберігають семантичну інформацію. Ці вектори [16] передаються до моделі XGBoost для подальшої класифікації. У цьому дослідженні запропоновано удосконалений підхід до класифікації текстових даних, що інтегрує потужний метод машинного навчання – XGBoost, з передовими техніками обробки природної мови.

Особливістю цього підходу стало вдосконалення можливостей моделі XGBoost через додавання методів інтерпретації, таких як SHAP (Shapley Additive Explanations), що дозволяє не лише підвищити точність моделі, а й забезпечити глибоке розуміння її рішень.

Ключовим удосконаленням є інтеграція технології SHAP, яка дозволяє з'ясувати, як окремі ознаки впливають на результати класифікації. Це дає можливість моделі не лише генерувати точні прогнози, але й відкривати внутрішню структуру прийняття рішень, що є важливим для задач, де інтерпретованість результатів є критично важливою. Завдяки використанню SHAP, модель стає прозорою і її результати можна пояснити з точки зору вкладу кожної ознаки в кінцеве рішення, що значно підвищує її надійність та застосовність у реальних умовах.

Зокрема, застосування SHAP дає можливість детально проаналізувати, чому певні текстові характеристики (наприклад, настрій, тональність або конкретні ключові слова) відіграють важливу роль у класифікації. Це дозволяє не лише підвищити точність класифікації, але й зменшити ймовірність помилок або непередбачених результатів, адже кожен етап прийняття рішення моделі стає більш зрозумілим.

Це удосконалення дозволяє зберігати високу продуктивність навіть при роботі з великими обсягами даних. Вдосконалена модель XGBoost демонструє значну стійкість до змін у даних та забезпечує високу точність класифікації навіть у складних ситуаціях, що дозволяє значно розширити сферу її застосування, зокрема в аналітиці текстів, прогнозуванні емоційного стану, або в інших дослідженнях, що потребують детального аналізу текстових даних.

Інтеграція SHAP в модель XGBoost забезпечує не тільки високу точність класифікації, але й ефективну інтерпретацію результатів, що дозволяє користувачам та дослідникам більш глибоко розуміти механізми роботи моделі та адаптувати її до специфічних вимог. Також модель показує високу продуктивність навіть при роботі з великими даними, роботу моделі зображено на рисунку 2.

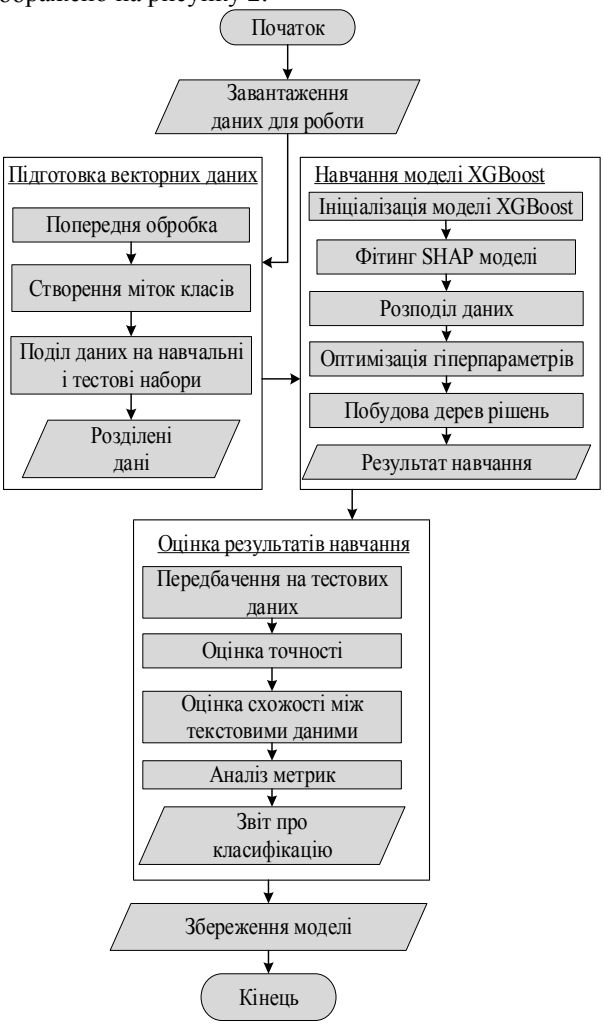

Рисунок 2 – Алгоритм роботи моделі XGBoost

DOI 10.15588/1607-3274-2025-2-14

Після тренування моделі необхідно оцінити її продуктивність на тестовому наборі. Для цього використовуються такі метрики, як точність, матриця плутанини, звіт про класифікацію.

Ці метрики дозволяють оцінити, наскільки добре модель класифікує повідомлення за їхнім настроєм, і визначити, чи модель може правильно передбачити позитивний, негативний або нейтральний настрій.

Для більш детального аналізу результатів використовується метрика Sentence Embedding Similarity, яка дозволяє вимірювати схожість між векторами повідомлень. Це корисно для задач, що вимагають не тільки класифікації, а й виявлення подібності між текстами, що можуть бути схожими за змістом, але мати різні настрої.

Такий підхід дозволяє здійснювати більш глибокий аналіз текстів і покращити якість класифікації. Блок-схема алгоритму роботи інтелектуальної системи зображена на рисунку 3.

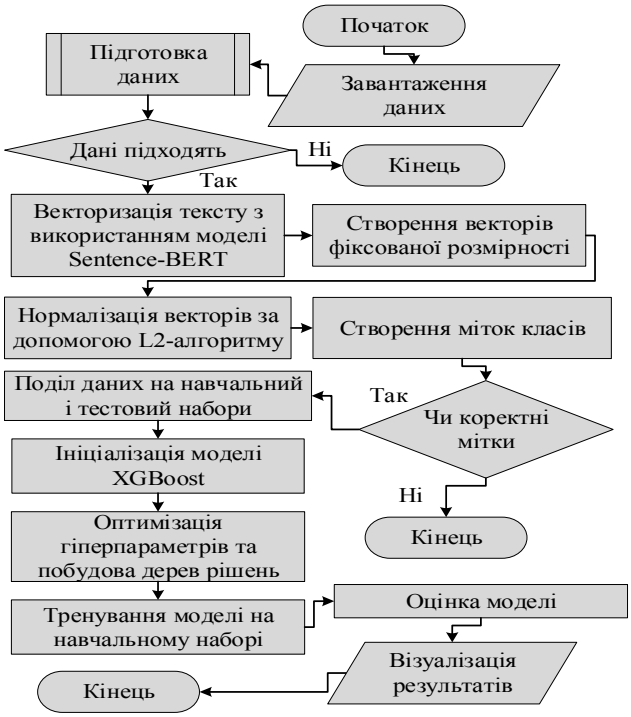

Рисунок 3 – Схема алгоритму інтелектуальної системи

Дослідження датасету Sentiment140 з використанням методів Sentence-BERT, XGBoost та Sentence Embedding Similarity дозволяє створити ефективну інформаційну систему для класифікації повідомлень за настроєм. Ключовими етапами є попередня обробка текстових даних, векторизація та нормалізація, навчання моделі на числових ознаках і подальша оцінка продуктивності за допомогою різних метрик. Важливою частиною є також аналіз схожості між повідомленнями, що дає можливість покращити якість класифікації і забезпечити високу точність моделі.

4 ЕКСПЕРИМЕНТИ

Для дослідження датасету Sentiment140 було обрано мову програмування Python 3.12 та інтегроване середовище розробки PyCharm. Початковим етапом є завантаження необхідного датасету, що містить дані з Twitter, де кожне користувацьке повідомлення супроводжується міткою, що вказує на його емоційний настрій. Датасет Sentiment140 включає такі характеристики: текст повідомлення (текстова ознака – повідомлення залишене в соціальній мережі в певний момент часу розміром до 280 символів), емоційний настрій повідомлення (цільова ознака), а також додаткові параметри, зокрема час публікації та інші метадані. Метою дослідження є вивчення взаємозв'язків між текстовими ознаками та емоційним настроєм в контексті соціальних медіа. Основною метою аналізу цих даних є побудова моделі машинного навчання, здатної класифікувати повідомлення за настроєм (позитивний, негативний або нейтральний) на основі текстового змісту.

Аналіз даних дозволить оцінити ефективність побудованої моделі та її здатність до точного передбачення емоційного настрою користувачів на основі текстових даних, що в свою чергу може бути корисним для аналізу настроїв у великих обсягах текстів в реальному часі. Завантажений датасет зображено на рисунку 4.

Рисунок 4 – Завантажений датасет

Оцінка точності моделі після навчання показала високі результати, що свідчить про її здатність ефективно класифікувати текстові дані. Модель досягла загальної точності 90% на тестовому наборі. Аналіз окремих метрик для кожного класу свідчить про збалансовану продуктивність.

Для класу negative модель продемонструвала точність 91%, повноту 87% та F1-метрику 89%. Це вказує на те, що модель добре ідентифікує негативні приклади, хоча частина з них була класифікована неправильно.

Для класу positive результати показали точність 89%, повноту 92% та F1-метрику 91%. Модель демонструє дещо кращу здатність виявляти позитивні класи, при цьому зберігаючи високу збалансованість між точністю та повнотою.

Зведені метрики показали середню макро-оцінку (macro avg) для точності, повноти та F1-метрики на рівні 90%, що свідчить про рівномірну продуктивність моделі для кожного класу та зважену середню оцінку (weighted avg), яка враховує частоту кожного класу в даних, також становила 90%,

DOI 10.15588/1607-3274-2025-2-14

підкреслюючи стабільну роботу моделі незалежно від кількості зразків кожного класу, дані про точність навчання, повноту та F1-метрику зображено на рисунку 5.

Точність навчання: 0.90

	precision	recall	f1-score	support
negative	0.90	0.90	0.90	239611
positive	0.90	0.90	0.90	240389
accuracy			0.90	480000
macro avg	0.90	0.90	0.90	480000
weighted avg	0.90	0.90	0.90	480000

Рисунок 5 – Інформація про створену модель нейронної мережі

Результати моделі класифікації продемонстрували високу точність і збалансованість у прогнозуванні класів negative і positive. AUC значення для ROC-кривої становить 0,90, що свідчить про високий рівень розділення між класами.

ROC-криві показують стабільну продуктивність моделі при зміні порогу класифікації. Precision-Recall аналіз демонструє для класу negative зростання точності (precision) при зменшенні повноти (recall), досягаючи максимуму в 1,0 для високих порогів.

Повнота зменшується із зниженням порогу, демонструючи стабільність на початкових рівнях і поступове зниження для менш релевантних зразків. Відповідно для класу positive, максимальна точність досягнута на рівні 1,0, що вказує на відмінне виявлення цього класу в певних умовах.

Повнота показує стабільне зменшення, залишаючись значною на середніх рівнях. Числові значення матриці плутанини та результати Precision-Recall аналізу зображено на рисунку 6, матрицю плутанини зображено на рисунку 7.

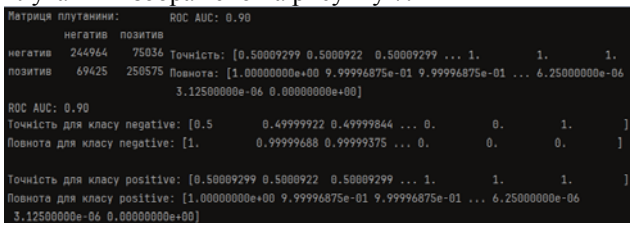

Рисунок 6 – Матриця плутанини та результати Precision-Recall аналізу

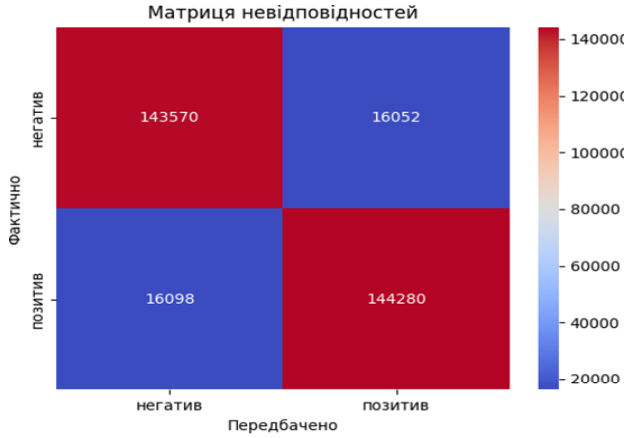

Рисунок 7 – Графічне відображення отриманої матриці плутанини

Ці результати підкреслюють високий рівень точності моделі при класифікації текстових даних, що особливо важливо для завдань аналізу настроїв, де правильна ідентифікація позитивних і негативних висловлювань має критичне значення. Модель показала здатність ефективно працювати з збалансованими наборами даних, що свідчить про її здатність обробляти різноманітні текстові дані без переваги одного класу над іншим, що є важливим для реальних застосувань. Візуалізація ROC-кривої на рисунку 8 чітко демонструє, як добре модель розрізняє твіти з різними настроями, що дозволяє швидко оцінити її ефективність.

Високе значення AUC (0,90) підтверджує, що модель володіє відмінною здатністю до класифікації з мінімальними хибними позитивними та негативними помилками. Площа під ROC-кривою є індикатором того, як добре модель розрізняє класи, і велика площа означає, що кількість хибних рішень значно зменшена.

ROC-крива на рисунку 8 ілюструє, наскільки добре модель розрізняє твіти з негативним і позитивним настроєм.

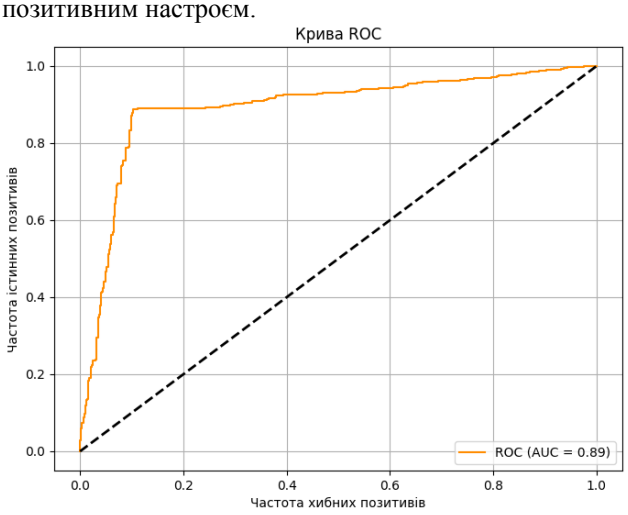

Рисунок 8 – ROC-крива поділу позитивних та негативних класів повідомлень користувачів

Precision-Recall крива на рисунку 9 демонструє, як змінюється співвідношення між точністю і повнотою для класифікації позитивних і негативних твітів залежно від порогового значення. Висока точність свідчить про те, що більшість твітів, які модель класифікує як позитивні або негативні, дійсно належать до відповідного класу. Це важливо для уникнення хибно-позитивних висновків. Висока повнота вказує, що модель здатна знайти всі твіти відповідного настрою в тестовій вибірці, що мінімізує хибно-негативні прогнози. Датасет Sentiment140 складається з коротких текстів (користувацьких повідомлень у Twitter). Такі дані часто містять емоційно насичені слова, скорочення, емодзі або інші нелінійні мовні конструкції, що ускладнює класифікацію. У цьому випадку Precision-Recall крива

DOI 10.15588/1607-3274-2025-2-14

демонструє, що модель ефективно працює навіть у таких умовах, забезпечуючи надійні прогнози.

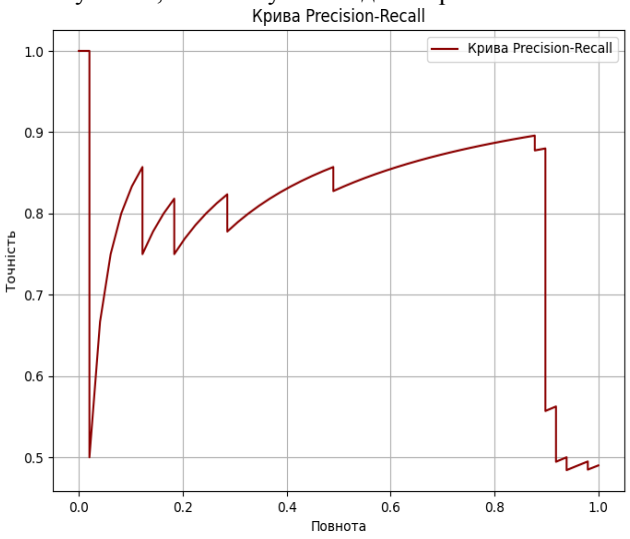

Рисунок 9 – Precision-Recall крива

Калібрувальна крива зображена на рисунку 10 є важливим інструментом для оцінки узгодженості між передбаченими ймовірностями моделі та фактичною часткою позитивних результатів у тестовій вибірці.

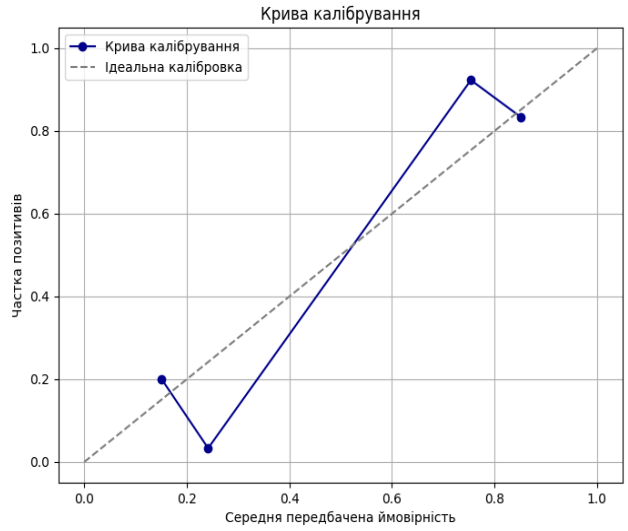

Рисунок 10 – Калібрувальна крива для оцінки узгодженості

5 РЕЗУЛЬТАТИ

Модель демонструє високу точність для передбачених ймовірностей у діапазоні [0,7579–0,8490], де фактична частка позитивних тісно корелює з передбаченнями 0,9286 і 0,8519 відповідно. Це вказує на надійність моделі для впевнених прогнозів.

У нижніх інтервалах ([0,1449–0,2499]) спостерігається недооцінка частки позитивних результатів, що може вказувати на необхідність подальшого калібрування моделі для покращення її продуктивності в цих діапазонах.

Косинусна схожість є одним із ключових інструментів для кількісного оцінювання семантичної подібності між текстовими даними. У даному дослідженні ми обчислили матрицю косинусної схожості між п'ятьма твітами з використанням векторизації Sentence-BERT. Значення косинусної схожості варіюються від 0,636 до 1,000, це зображено на рисунках 11 та 12.

Найвищі показники свідчать про високу семантичну схожість між відповідними текстами, що може вказувати на схожість тематики або ключових виразів у цих твітах.

Результати матриці косинусної схожості підтверджують ефективність Sentence-BERT у виявленні семантичної подібності між текстами. Високі значення для певних пар твітів (наприклад, Твіт 3 і Твіт 4) свідчать про те, що модель правильно ідентифікує близькість змісту, навіть якщо тексти мають відмінності на рівні синтаксису або поверхневої структури.

```
Матриця косинусної схожості між першими 5 твітами:
          Твіт 1    Твіт 2    Твіт 3    Твіт 4    Твіт 5
Твіт 1  1.000000  0.760375  0.772464  0.720631  0.655470
Твіт 2  0.760375  1.000000  0.726243  0.760966  0.748869
Твіт 3  0.772464  0.726243  1.000000  0.650697  0.690497
Твіт 4  0.720631  0.760966  0.650697  1.000000  0.697242
Твіт 5  0.655470  0.748869  0.690497  0.697242  1.000000
Дані калібрувальної кривої:
Середні передбачені ймовірності: [0.15001779 0.2499676  0.75009756 0.85007704]
Частка позитивних: [0.10068988 0.09972455 0.89976803 0.89893478]
```

Рисунок 11 – Матриця косинусної схожості між першими 5 твітами

Нижчі значення (наприклад, Твіт 2 і Твіт 5) можуть вказувати на тексти, які менш пов'язані тематично або мають суттєві відмінності у виразах, що є важливим для аналізу різноманітності даних.

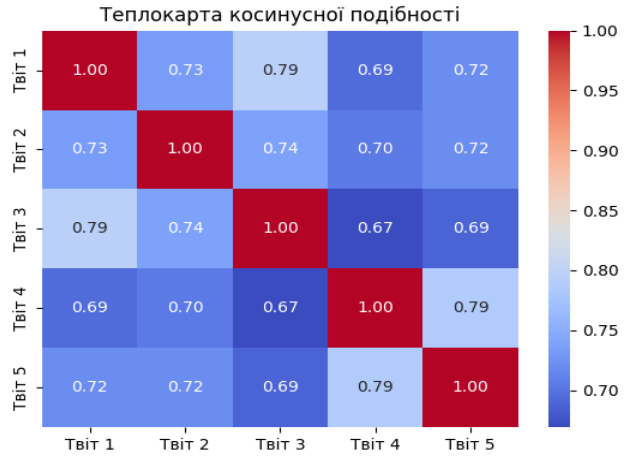

Рисунок 12 – Теплова карта косинусної схожості між першими 5 твітами

F1-міра є важливим показником у задачах класифікації, що об'єднує як точність, так і повноту у єдину метрику, забезпечуючи комплексне розуміння продуктивності моделі, що зображено на рисунку 13. У цьому дослідженні було оцінено F1-міру для різних порогових значень у діапазоні від 0 до 1 з кроком 0,0204. Найвища F1-міра, 0,9074, досягнута для порогів від 0,306 до 0,673, вказує на ефективну продуктивність моделі в цьому діапазоні. Це

DOI 10.15588/1607-3274-2025-2-14

підтверджує здатність моделі точно ідентифікувати як позитивні, так і негативні класи в умовах збалансованого співвідношення точності й повноти. Збереження високого значення F1-міри у середньому діапазоні порогів (0,306–0,673) свідчить про високу стійкість моделі до змін порогових значень, що є важливим аспектом її надійності.

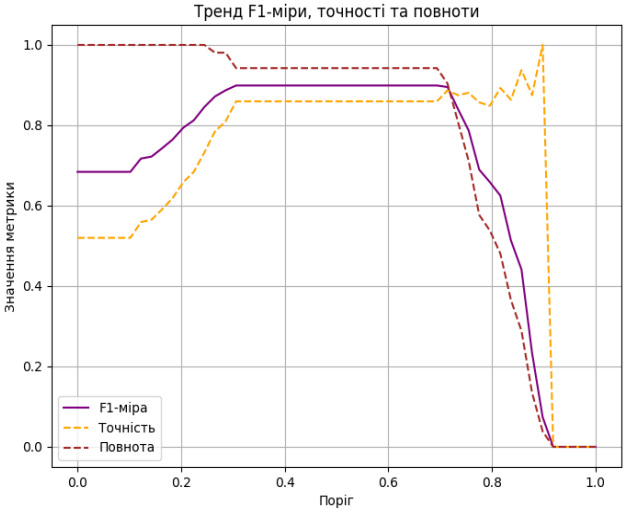

Рисунок 13 – Тренди F1-міри, точності та повноти

Кумулятивний приріст є ключовою характеристикою, що дозволяє оцінити прогресивну зміну продуктивності моделі при аналізі класифікаційних задач.

У цьому дослідженні на рисунку 14 було розглянуто послідовне накопичення правильно класифікованих прикладів для визначення динаміки ефективності моделі.

Стабільний лінійний ріст у середньому діапазоні підкреслює послідовність та надійність моделі, а формування плато у верхньому діапазоні сигналізує про досягнення межі можливостей моделі, що є типовим для задач із високим ступенем складності.

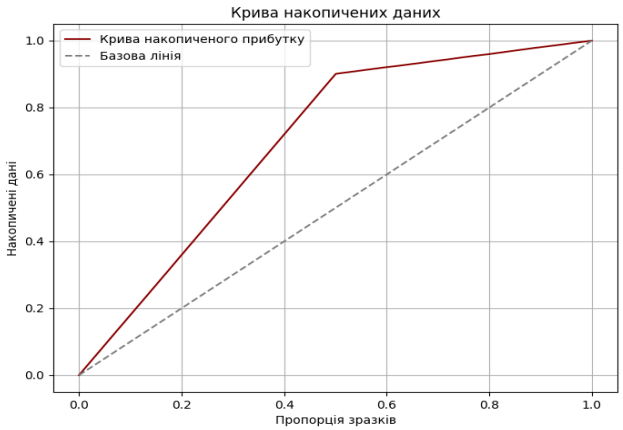

Рисунок 14 – Кумулятивна крива приросту

Отримані результати свідчать про високу продуктивність моделі на основі аналізу різних метрик: матриця плутанини демонструє точність класифікації між класами, ROC-крива підтверджує баланс між чутливістю та специфічністю, а Precision-Recall крива вказує на здатність моделі ефективно працювати з незбалансованими даними.

Калібрувальна крива відображає відповідність передбачених ймовірностей фактичним результатам, тоді як матриця косинусної схожості ілюструє семантичну близькість текстів. Тренди F1-міри, точності та повноти підкреслюють стабільність моделі на різних етапах роботи, що робить її придатною для широкого спектра завдань класифікації.

6 ОБГОВОРЕННЯ

Було здійснено порівняння трьох різних підходів до аналізу текстових даних, кожен з яких використовує комбінації різних методів обробки та моделювання: TF-IDF + onehotencoding + cosine similarity, BERT + XGBoost + sentence embedding similarity, та S-BERT + XGBoost + SHAP + sentence embedding similarity + lambda.

Перший підхід, TF-IDF + onehotencoding + cosine similarity, є класичним методом для текстової класифікації, який на основі частоти термінів в документі створює прості векторні представлення слів, що дозволяє оцінити подібність між текстами за допомогою косинусної міри.

Незважаючи на свою простоту і ефективність для невеликих наборів даних, цей підхід має значні обмеження, оскільки не враховує контексту слів та їх взаємозв'язків у тексті.

Крім того, методи onehotencoding та TF-IDF створюють великі векторні простори для великих наборів даних, що може призвести до зниження продуктивності моделі, особливо коли йдеться про обробку складніших мовних конструкцій.

Другий підхід, який включає BERT + XGBoost + sentence embedding similarity, використовує передові методи трансформерів для представлення текстів у вигляді векторів, що зберігають контекстуальну інформацію слів. BERT значно покращує результат, оскільки здатен усувати багато недоліків класичних методів векторизації, таких як TF-IDF, враховуючи семантичні зв'язки між словами та їхніми контекстами. Використання XGBoost як моделі для класифікації дозволяє ефективно обробляти дані, знижуючи ризик перенавчання завдяки вбудованим методам регуляризації. Проте, цей підхід все ще має певні обмеження при обробці великих обсягів даних, оскільки використання BERT для генерації векторів потребує значних обчислювальних ресурсів, що може бути проблематичним при масштабуванні. На рисунку 15 зображено порівняння часу виконання моделей, на рисунку 16 зображено порівняння використання пам'яті моделей.

DOI 10.15588/1607-3274-2025-2-14

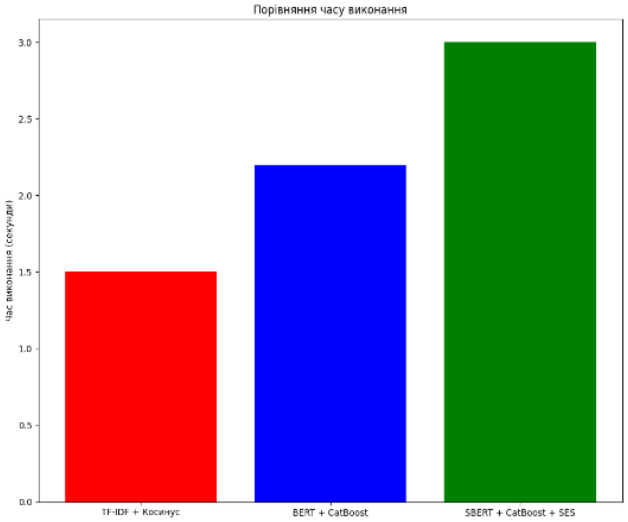

Рисунок 15 – Порівняння часу виконання

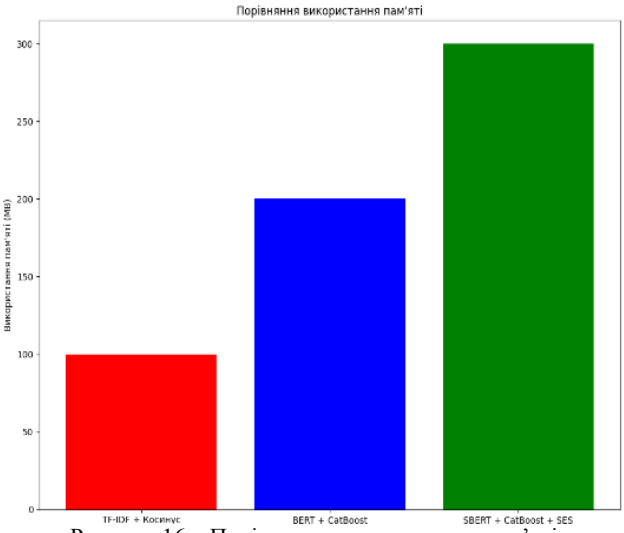

Рисунок 16 – Порівняння використання пам’яті

Останній підхід, S-BERT + XGBoost + SHAP + sentence embedding similarity + lambda, представляє собою більш складну та адаптовану до реальних умов модель. S-BERT (Sentence-BERT) використовує модифікацію BERT, яка дозволяє ефективно працювати з парами текстів, що дає змогу швидше обробляти дані та генерувати високоякісні векторні представлення для задач порівняння текстів. Використання XGBoost з доповненням SHAP дозволяє отримати прозорість моделі через оцінку важливості ознак, що є важливим для розуміння, як модель приймає свої рішення.

Крім того, інтеграція методу lambda допомагає в адаптації до великих обсягів даних і підвищує загальну продуктивність моделі за рахунок вдосконаленого налаштування параметрів та гнучкості в обробці текстів. Цей підхід дає значно кращу точність порівняно з іншими через його здатність ефективно працювати з великими обсягами даних, мінімізуючи час обробки та зберігаючи високу точність класифікації. На рисунку 17 зображено порівняння часу розміру моделей, на рисунку 18 зображено порівняння точності моделей.

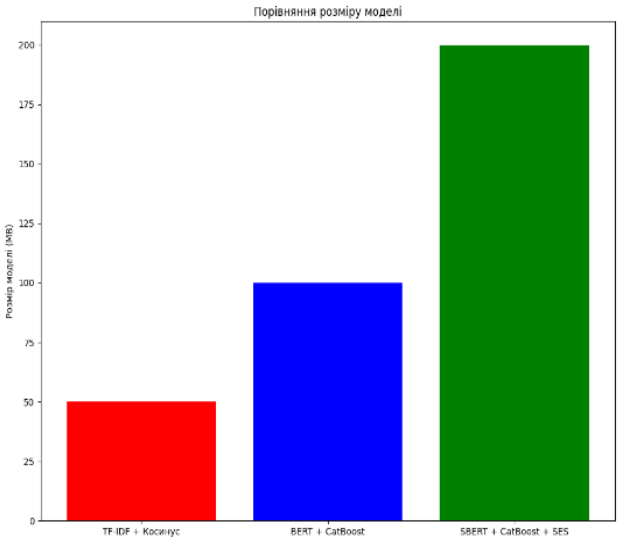

Рисунок 17 – Порівняння розміру моделі

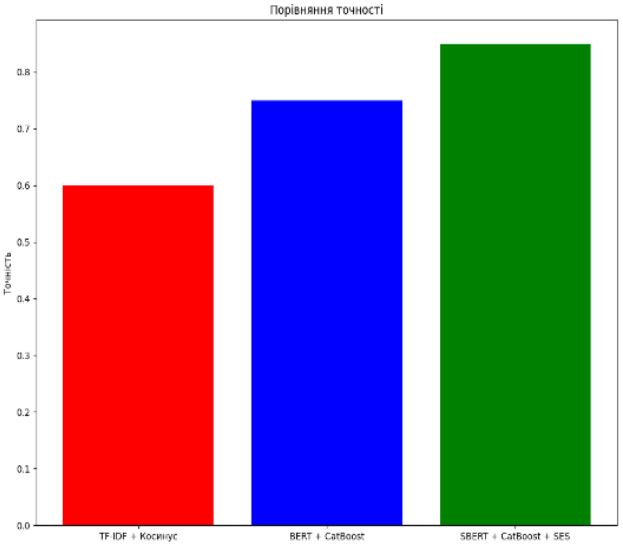

Рисунок 18 – Порівняння точності

Основною перевагою підходу S-BERT + XGBoost + SHAP + sentence embedding similarity + lambda є його здатність до масштабування і гнучкості. Завдяки використанню S-BERT, який є швидшою модифікацією BERT для задач порівняння текстів, і SHAP для інтерпретації результатів, цей підхід може бути використаний для обробки великих наборів даних без втрати точності. Це робить його ідеальним для застосувань, де потрібно обробляти великий обсяг текстової інформації та забезпечувати не тільки високу точність, але й прозорість роботи моделі.

Отже, S-BERT + XGBoost + SHAP + sentence embedding similarity + lambda є найефективнішим підходом серед трьох, здатним забезпечити найкращу точність на великих наборах даних. Це свідчить про важливість використання глибших і більш адаптованих до специфічних задач моделей у

DOI 10.15588/1607-3274-2025-2-14

поєднанні з методами інтерпретації результатів для забезпечення надійних і точних результатів.

ВИСНОВКИ

У даному дослідженні вирішено проблему ефективного аналізу текстових даних у задачах класифікації емоційного забарвлення текстів, що є критично важливим у сучасних умовах постійного зростання обсягу інформації в соціальних мережах. Запропонована методологія поєднує можливості трансформерної архітектури Sentence-BERT для отримання контекстуально збагачених векторних представлень текстів із потужністю градієнтного бустингу XGBoost, що забезпечує високу точність та узагальнюючу здатність класифікації.

Додатково було застосовано метод SHAP для пояснення прийнятих модельних рішень, що дозволяє оцінити вплив окремих ознак на результати передбачень та підвищити довіру до моделі у практичних застосуваннях. Крім того, використання метрики Sentence Embedding Similarity дозволило ефективно виявляти приховані семантичні зв'язки між текстами, що відкриває можливості для глибшого аналізу контенту.

Наукова новизна отриманих результатів полягає в удосконаленні методів текстової класифікації шляхом комбінування методів глибокого навчання та ансамблевих алгоритмів. Вперше продемонстровано, що інтеграція Sentence-BERT та XGBoost із застосуванням інтерпретаційних методів дозволяє не лише підвищити точність класифікації текстів за емоційним забарвленням, але й зробити модель більш прозорою для користувачів. Це особливо важливо для критичних застосувань, таких як аналіз громадської думки, автоматизована модерація контенту та системи підтримки ухвалення рішень, де автоматичне пояснення моделі є ключовим фактором.

Використання Sentence-BERT у поєднанні з класифікатором XGBoost також є перспективним для побудови інтелектуальних систем рекомендацій, де аналіз емоційного забарвлення відгуків може допомагати у формуванні персоналізованих пропозицій.

Запропонований підхід продемонстрував високу точність і стабільність результатів навіть при аналізі великих масивів даних, що свідчить про його масштабованість та придатність для використання в реальних умовах. Одним із важливих аспектів дослідження є оцінка роботи моделі на різних порогових значеннях класифікації, що дало змогу досягти оптимального балансу між точністю та повнотою, забезпечуючи гнучкість у налаштуванні моделі залежно від конкретних завдань.

Практична цінність роботи полягає у можливості використання запропонованої методології для широкого спектра задач, що вимагають аналізу тональності текстів. Зокрема, вона може бути ефективно застосована у сфері аналізу соціальних мереж для оцінки настроїв користувачів та виявлення тенденцій у суспільних дискусіях. Важливим напрямком подальших досліджень є також адаптація моделі до роботи з незбалансованими наборами даних, що є поширеною проблемою у багатьох реальних сценаріях аналізу текстової інформації.

ЛІТЕРАТУРА

1. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering / [M. Mujahid, E. Kına, F. Rustam et al.] // Journal of Big Data. – 2024. – Vol. 11, No. 1. – P. 1–32. DOI: 10.1186/s40537-024-00943-4
2. Neural network signal integration from thermogas-dynamic parameter sensors for helicopters turboshaft engines at flight operation conditions / [S. Vladov, L. Scislo, V. Sokurenko et al.] // Sensors. – 2024. – Vol. 24, No. 13. – P. 4246. DOI: 10.3390/s24134246
3. Batiuk T. Intellectual system for clustering users of social networks derived from the message sentiment analysis / T. Batiuk, D. Dosyn // Journal of Lviv Polytechnic National University "Information Systems and Networks". – 2023. – Vol. 13. – P. 121–138. DOI: 10.23939/sisn2023.13.121
4. Ensemble learning with pre-trained transformers for crash severity classification: a deep N.L.P. approach / [S. Jaradat, R. Nayak, A. Paz, M. Elhenawy] // Algorithms. – 2024. – Vol. 17, No. 7. – P. 284. DOI: 10.3390/a17070284
5. The method of restoring lost information from sensors based on auto-associative neural networks / S. Vladov, R. Yakovliev, V. Vysotska et al.] // Applied System Innovation. – 2024. – Vol. 7, no. 3. – P. 53. DOI: 10.3390/asi7030053
6. Core technology topic identification and evolution analysis based on patent text mining – a case study of unmanned ship / [Y. Lin, X. Wang, J. Yang, S. Wang] // Applied Sciences. – 2024. – Vol. 14, No. 11. – P. 4661. DOI: 10.3390/app14114661
7. The impact of input types on smart contract vulnerability detection performance based on deep learning: a preliminary study / [I. M. Aldyaflah, W. Zhao, S. Yang, X. Luo] // Information. – 2024. – Vol. 15, no. 6. – P. 302. DOI: 10.3390/info15060302
8. Batiuk T. A realization of visual biometric validation to enhance guarded and efficient authorization for intellectual systems / T. Batiuk, D. Dosyn // CEUR Workshop Proceedings, 8th Intern. Conf. on Computational Linguistics and Intelligent Systems COLINS 2024. – 2024. – Vol. 3668. – P. 247–268.
9. Ivokhin E. Restructuring of the model "State–Probability of Choice" based on products of stochastic rectangular matrices / E. Ivokhin, O. Oletsky // Cybernetics and Systems Analysis. – 2022. – Vol. 58, No. 2. – P. 242–250. DOI: 10.1007/s10559-022-00456-z
10. Information technology for the operational processing of military content for commanders of tactical army units / [V. Danylyk, V. Vysotska, V. Andrunyk et al.] // International Journal of Computer Network and

DOI 10.15588/1607-3274-2025-2-14

Information Security. – 2024. – Vol. 16, No. 3. – P. 115–143. DOI: 10.5815/ijcnis.2024.03.09
11. Batiuk T. Realization of the decision-making support system for Twitter users' publications analysis / T. Batiuk, D. Dosyn // Radio Electronics Computer Science Control. – 2024. – Vol. 1, No. 24. – P. 175–187. DOI: 10.15588/1607-3274-2024-1-16
12. Oletsky O. Exploring dynamic equilibrium of alternatives on the base of rectangular stochastic matrices / O. Oletsky // CEUR Workshop Proceedings, Modern Machine Learning Technologies and Data Science Workshop MoMLeT&DS 2021. – 2021. – Vol. 2917. – P. 151–160.
13. Oletsky O. On constructing adjustable procedures for enhancing consistency of pairwise comparisons on the base of linear equations / O. Oletsky // CEUR Workshop Proceedings. – 2021. – Vol. 3106. – P. 177–185.
14. Lin Y. Enhanced Transformer-BLSTM model for classifying sentiment of user comments on movies and books / Y. Lin, T. Liu // IEEE Access. – 2024. – P. 1–1. DOI: 10.1109/access.2024.3416755
15. Intellectual system for socialization of individuals with contributed interests derived from NLP, machine learning, and SEO algorithms / [T. Batiuk, V. Vysotska, R. Holoshchuk, S. Holoshchuk] // CEUR Workshop Proceedings, 6th Intern. Conf. on Computational Linguistics and Intellectual Systems COLINS 2022. – 2022. – Vol. 3171. – P. 572–631.
16. Oletsky O. A model of information influences on the base of rectangular stochastic matrices in chains of reasoning with possible contradictions / O. Oletsky // CEUR Workshop Proceedings, IT&I Workshops 2021. – 2021. – Vol. 3179. – P. 354–361.

Accepted 05.02.2025.
Received 29.04.2025.

UDC 004.9

NATURAL LANGUAGE PROCESSING OF SOCIAL MEDIA TEXT DATA USING BERT AND XGBOOST

Batiuk T. – Post-graduate student of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.
Dosyn D. – Doctor of Sciences, Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context The growth of text data in social networks requires the development of effective methods for sentiment analysis that can take into account both lexical and contextual dependencies. Traditional approaches to text processing have limitations in understanding semantic relationships between words, which affects the accuracy of classification. The integration of deep neural networks for text vectorization with ensemble machine learning algorithms and methods for interpreting results allows improving the quality of sentiment analysis.

Objective. The aim of the study is to develop and evaluate a new approach to text message sentiment classification that combines Sentence-BERT for deep semantic vectorization, XGBoost for high-accuracy classification, SHAP for explaining the contribution of features, sentence embedding similarity for assessing semantic similarity, and λ-regularization to improve the generalization ability of the model. The study is aimed at analyzing the impact of these methods on the quality of classification, identifying the most significant features and optimizing parameters.

Method. The study uses Sentence-BERT to transform text data into a vector space with deep semantic connections. XGBoost is used for sentiment classification, which provides high accuracy and stability even on unevenly distributed datasets. The SHAP method is used to explain the contribution of features, which allows us to determine which factors have the greatest impact on the prediction. Additionally, sentence embedding similarity is used to compare texts.

Results. The proposed approach demonstrates high efficiency in mood classification tasks. The ROC-AUC value confirms the ability of the model to accurately distinguish between classes of emotional coloring of the text. The use of SHAP ensures the interpretability of the results, allowing us to explain the influence of each feature on the classification. Sentence embedding similarity confirms the efficiency of Sentence-BERT in detecting semantically similar texts, and λ-regularization improves the generalization ability of the model.

Conclusions. The study demonstrates scientific novelty through a comprehensive combination of Sentence-BERT, XGBoost, SHAP, sentence embedding similarity, and λ-regularization to improve the accuracy and interpretability of sentiment analysis. The results obtained confirm the effectiveness of the proposed approach, which makes it promising for application in public opinion monitoring, automated content moderation, and personalized recommendation systems. Further research can be aimed at adapting the model to specific domains and improving interpretation methods.

KEYWORDS: Machine learning, feature normalization, Transformers, confusion matrix, Sentence-BERT, text data classification.

REFERENCES

1. Mujahid M. Kına E., Rustam F., Villar M. G., Alvarado E. S., Diez I. D. L. T., Ashraf I. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering, *Journal of Big Data*, 2024,

DOI 10.15588/1607-3274-2025-2-14

Vol. 11, No. 1, pp. 1–32. DOI: 10.1186/s40537-024-00943-4
2. Vladov S., Scislo L., Sokurenko V., Muzychuk O., Vysotska V., Osadchy S., Sachenko A. Neural network signal integration from thermogas-dynamic parameter sensors for helicopters turboshaft engines at flight operation conditions, *Sensors*, 2024, Vol. 24, No. 13, P. 4246. DOI: 10.3390/s24134246
3. Batiuk T., Dosyn D. Intellectual system for clustering users of social networks derived from the message sentiment analysis, *Journal of Lviv Polytechnic National University* "Information Systems and Networks", 2023, Vol. 13, pp. 121–138. DOI: 10.23939/sisn2023.13.121
4. Jaradat S., Nayak R., Paz A., Elhenawy M. Ensemble learning with pre-trained transformers for crash severity classification: a deep N.L.P. approach, *Algorithms,* 2024, Vol. 17, No. 7, P. 284. DOI: 10.3390/a17070284
5. Vladov S., Yakovliev R., Vysotska V., Nazarkevych M., Lytvyn V. The method of restoring lost information from sensors based on auto-associative neural networks, *Applied System Innovation*, 2024, Vol. 7, No. 3, P. 53. DOI: 10.3390/asi7030053
6. Lin Y., Wang X., Yang J., Wang S. Core technology topic identification and evolution analysis based on patent text mining – a case study of unmanned ship, A*pplied Sciences*, 2024, Vol. 14, No. 11, P. 4661. DOI: 10.3390/app14114661
7. Aldyaflah I., Zhao W., Yang S., Luo X. The impact of input types on smart contract vulnerability detection performance based on deep learning: a preliminary study, *Information*, 2024, Vol. 15, No. 6, P. 302. DOI: 10.3390/info15060302
8. Batiuk T., Dosyn D. A realization of visual biometric validation to enhance guarded and efficient authorization for intellectual systems, *CEUR Workshop Proceedings, 8th Intern. Conf. on Computational Linguistics and Intelligent Systems COLINS 2024*, 2024, Vol. 3668, pp. 247–268.
9. Ivokhin E., Oletsky O. Restructuring of the model "State–Probability of Choice" based on products of stochastic rectangular matrices, *Cybernetics and Systems Analysis,* 2022, Vol. 58, No. 2, pp. 242–250. DOI: 10.1007/s10559-022-00456-z
10. Danylyk V., Vysotska V., Andrunyk V., Uhryn D., Ushenko Y. Information technology for the operational processing of military content for commanders of tactical army units, *International Journal of Computer Network and Information Security*, 2024, Vol. 16, No. 3, pp. 115–143. DOI: 10.5815/ijcnis.2024.03.09
11. Batiuk T., Dosyn D. Realization of the decision-making support system for Twitter users' publications analysis, *Radio Electronics Computer Science Control,* 2024, Vol. 1, No. 24, pp. 175–187. DOI: 10.15588/1607-3274-2024-1-16
12. Oletsky O. Exploring dynamic equilibrium of alternatives on the base of rectangular stochastic matrices, *CEUR Workshop Proceedings, Modern Machine Learning Technologies and Data Science Workshop MoMLeT&DS* 2021, 2021, Vol. 2917, pp. 151–160.
13. Oletsky O. On constructing adjustable procedures for enhancing consistency of pairwise comparisons on the base of linear equations, *CEUR Workshop Proceedings,* 2021, Vol. 3106, pp. 177–185.
14. Lin Y., Liu T. Enhanced Transformer-BLSTM model for classifying sentiment of user comments on movies and books, *IEEE Access,* 2024, pp. 1–1. DOI: 10.1109/access.2024.3416755
15. Batiuk T., Vysotska V., Holoshchuk R., Holoshchuk S. Intellectual system for socialization of individuals with contributed interests derived from NLP, machine learning, and SEO algorithms, *CEUR Workshop Proceedings, 6th Intern. Conf. on Computational Linguistics and Intellectual Systems COLINS* 2022, 2022, Vol. 3171, pp. 572–631.
16. Oletsky O. A model of information influences on the base of rectangular stochastic matrices in chains of reasoning with possible contradictions, *CEUR Workshop Proceedings, IT&I Workshops 2021,* 2021, Vol. 3179, pp. 354–361.

DOI 10.15588/1607-3274-2025-2-14