Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра мультимедійних систем

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: «**АНАЛІЗ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ПОВІДОМЛЕНЬ У СОЦІАЛЬНИХ МЕРЕЖАХ**»

Виконала: студенка 4-го року навчання,

Освітньої програми «Інженерія програмного забезпечення», 121

Крячко Ірина Геннадіївна

Керівник Олецький О.В.,

Доцент, кандидат техн. наук

Рецензент

_____

(прізвище та ініціали)

Кваліфікаційна робота захищена з оцінкою

_____

Секретар ЕК

_____

«____» _____

20____ р.

Київ – 2025

Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра мультимедійних систем

ЗАТВЕРДЖУЮ

Зав.кафедри мультимедійних систем,

Доцент, кандидат наук

_____ Жежерун О.П.

(підпис)

“_____”_____ 2025

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

для кваліфікаційної роботи

студентці 4-го курсу, факультету інформатики

Крячко Ірині Геннадіївні

Тема: «Аналіз алгоритмів машинного навчання для класифікації повідомлень у соціальних мережах»

Зміст кваліфікаційної роботи:

Перелік прийнятих скорочень

Вступ

1. Теоретичні основи класифікації текстів у соціальних мережах
2. Аналіз предметної області і постановка задачі
3. Реалізація підходів на основі класичних алгоритмів
4. Реалізація підходів на основі сучасних алгоритмів
5. Порівняння методів тематичної класифікації

Висновки

Список літератури

Дата видачі “___”_____2025 Керівник _____

(підпис)

Завдання отримав _____

(підпис)

ГРАФІК ПІДГОТОВКИ КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

№ з/п	Перелік робіт	Термін виконання	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Отримання завдання на кваліфікаційну роботу	30.09.2024			
2.	Огляд технічної літератури за темою роботи	21.10.2024 - 27.01.2025			
3.	Аналіз специфіки текстових даних із соціальних мереж	03.02.2025 - 11.02.2025			
4.	Формування вимог до системи тематичної класифікації	12.02.2025 - 15.02.2025			
5.	Збір, обробка та очищення корпусу текстових повідомлень	17.02.2025 - 28.02.2025			
6.	Реалізація класичних методів	01.03.2025 - 24.03.2025			
7.	Реалізація сучасних методів	25.03.2025 - 17.04.2025			

8.	Порівняльний аналіз точності, інтерпретованості та гнучкості методів	18.04.2025 - 26.04.2025			
9.	Розробка рекомендацій щодо вибору алгоритмів тематичної класифікації	27.04.2025 - 08.05.2025			
10.	Попередній захист кваліфікаційної роботи	19.05.2025			
11.	Остаточне оформлення пояснювальної роботи та слайдів	20.05.2025 - 27.05.2025			
12.	Захист кваліфікаційної роботи	03.06.2025			

Графік узгоджено «____» _____ 2025р.

Науковий керівник Олецький Олексій Віталійович

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ 7
ВСТУП 8
РОЗДІЛ 1. Теоретичні основи класифікації текстів у соціальних мережах 10
1.1. Характеристика соціальних мереж як джерела текстових даних ... 10
1.2. Проблеми та особливості обробки текстів у соціальних медіа 11
1.3. Методи тематичної класифікації текстових повідомлень 12
РОЗДІЛ 2. Аналіз предметної області і постановка задачі 16
2.1. Аналіз типів повідомлень у соціальних мережах за темами 16
2.1.1. Обґрунтування вибору тематичних категорій 17
2.2. Постановка задачі тематичної класифікації повідомлень 17
2.3. Вимоги до систем класифікації 18
2.3.1. Функціональні вимоги 18
2.3.2. Нефункціональні вимоги 19
2.4. Обґрунтування вибору метрик якості класифікації 19
2.5. Обґрунтування вибору метрик якості кластеризації 22
РОЗДІЛ 3. Реалізація підходів на основі класичних алгоритмів машинного навчання 25
3.1. Опис та структура датасету 25
3.2. Векторизація тексту 27
3.3. Модель 1: TF-IDF + Naive Bayes 28
3.4. Модель 2: TF-IDF + Logistic Regression 31
3.5. Модель 3: TF-IDF + SVM 34
РОЗДІЛ 4. Реалізація підходів на основі сучасних алгоритмів 38
4.1. Тематичне моделювання за допомогою LDA 38
4.2. Семантичне групування повідомлень за допомогою BERT + кластеризації 40
4.2.1. Кластеризація за допомогою BERT + K-Means 41
4.2.2. Семантичне групування повідомлень за допомогою BERT + HDBSCAN 43
4.3. Zero-shot класифікація 45
РОЗДІЛ 5. Порівняння методів тематичної класифікації та оптимальні сценарії їх застосування 48
5.1. Порівняння класичних моделей машинного навчання 48

5.2. Порівняння нейромережевих та embedding-підходів 50
5.3. Аналіз zero-shot класифікації 52
5.4. Порівняльна характеристика методів тематичної класифікації 53
5.5. Рекомендації щодо вибору моделі 54
ВИСНОВКИ 58
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 60

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

NLP – Natural Language Processing (обробка природної мови);

TF-IDF – Term Frequency-Inverse Document Frequency;

LSTM – Long Short-Term Memory (рекурентна нейронна мережа довготривалої короткочасної пам’яті);

GRU – Gated Recurrent Unit (рекурентна нейронна мережа з механізмом ґейтування);

BERT – Bidirectional Encoder Representations from Transformers (двонаправлені представлення на основі трансформерів);

RoBERTa – Robustly Optimized BERT Approach (оптимізований варіант моделі BERT);

LDA – Latent Dirichlet Allocation (латентне розподілення Діріхле);

HDBSCAN – Hierarchical Density-Based Spatial Clustering of Applications with Noise (ієрархічна кластеризація на основі щільності з обробкою шумів).

SVM – Support Vector Machine (метод опорних векторів);

UMAP – Uniform Manifold Approximation and Projection (уніфіковане апроксимування та проєктування многовидів);

t-SNE – t-distributed Stochastic Neighbor Embedding (стохастичне вкладення сусідів з t-розподілом);

XLM-RoBERTa – Cross-lingual RoBERTa (багатомовна версія моделі RoBERTa, натренована на багатьох мовах);

CPU – Central Processing Unit (центральний процесор);

GPU – Graphics Processing Unit (графічний процесор).

ВСТУП

У сучасному цифровому світі соціальні мережі стали невід'ємною частиною повсякденного життя мільйонів людей. Щодня користувачі створюють величезний обсяг текстового контенту, який охоплює широкий спектр тем — від особистих думок і новин до політичних дискусій та рекламних повідомлень. Цей потік текстових даних є цінним ресурсом для аналізу суспільних тенденцій, однак водночас створює виклик — як автоматично й ефективно класифікувати такі повідомлення за тематикою, враховуючи їхню неструктурованість, емоційність, варіативність мови та обмежений контекст.

Тематична класифікація повідомлень у соціальних мережах є актуальним завданням в галузі обробки природної мови (NLP). Класичні підходи, як-от логістична регресія у поєднанні з TF-IDF, продемонстрували певну ефективність, однак мають обмеження щодо розуміння контексту. У відповідь на це сучасна наука звернулася до нейромережевих підходів: рекурентні нейронні мережі (LSTM, GRU), трансформери (BERT, RoBERTa), а також zero-shot класифікація стали провідними напрямами досліджень.

Попри значний науковий прогрес, актуальними залишаються практичні аспекти застосування нейромереж у задачах тематичної класифікації повідомлень, особливо у випадках багатотемності, відсутності розмічених даних або необхідності адаптації до нових тем. У зв’язку з цим виникає необхідність системного аналізу ефективності різних підходів на реальних даних із соціальних мереж.

Метою цього дослідження є визначити ефективність різних нейромережевих підходів для автоматичної класифікації повідомлень у соціальних мережах за тематикою.

Для досягнення поставленої мети передбачається виконати такі завдання:

- Проаналізувати особливості тематичної класифікації повідомлень у соціальних мережах та сформувати вимоги до системи класифікації.

- Реалізувати тпідходи на основі класичних методів, LDA, BERT із кластеризацією (k-means, HDBSCAN), а також zero-shot класифікацію.
- Провести експериментальні дослідження та оцінити якість кожної моделі за ключовими метриками.
- Виокремити переваги, обмеження та потенційні сфери застосування кожного підходу.
- Окреслити рекомендації щодо вибору моделі залежно від поставленого завдання класифікації.

Об’єктом дослідження є процес автоматичної тематичної класифікації текстових повідомлень у соціальних мережах.

Робота складається з п’яти розділів.

У першому розділі проаналізовано особливості текстів соціальних мереж та підходи до їх тематичної класифікації. Надано огляд класичних і сучасних методів обробки текстових даних, а також сформульовано постановку задачі класифікації.

У другому розділі здійснено аналіз типових тем соціальних повідомлень, визначено вимоги до систем класифікації, обґрунтовано вибір метрик для оцінки якості класифікації та кластеризації.

У третьому розділі реалізовано класичні підходи до тематичної класифікації на основі TF-IDF у поєднанні з алгоритмами Naive Bayes, Logistic Regression та SVM. Проведено оцінювання якості та порівняння результатів.

У четвертому розділі досліджено сучасні методи: тематичне моделювання з використанням LDA, кластеризацію на основі BERT-ембедінгів (K-Means і HDBSCAN), а також zero-shot класифікацію без навчання на специфічних класах.

П’ятий розділ присвячено порівнянню всіх підходів, їх характеристикам, перевагам і обмеженням, а також формуванню практичних рекомендацій щодо вибору оптимальної моделі для різних сценаріїв застосування.

РОЗДІЛ 1. Теоретичні основи класифікації текстів у соціальних мережах

У цьому розділі розглянуто теоретичні засади автоматичної класифікації текстів у соціальних мережах — однієї з найактуальніших задач сучасної обробки природної мови. Швидкий ріст обсягів користувацького контенту на таких платформах, як Twitter, Facebook, Reddit та Instagram, зумовлює необхідність ефективних інструментів для аналізу та тематичного структурування текстової інформації. Водночас, неформальний стиль повідомлень, наявність сленгу, емодзі та культурних контекстів створюють суттєві виклики для алгоритмів автоматичної обробки.

У межах розділу розглядаються такі ключові аспекти: особливості соціальних мереж як джерела текстових даних, проблеми, пов’язані з обробкою текстів у соціальних медіа, а також основні методи тематичної класифікації повідомлень.

Матеріал цього розділу створює концептуальне підґрунтя для подальшої реалізації та порівняння різних підходів у практичній частині дослідження.

1.1. Характеристика соціальних мереж як джерела текстових даних

Соціальні мережі стали одним із провідних джерел генерації текстового контенту у XXI столітті. Такі платформи, як Facebook, Twitter (нині X), Instagram, Reddit, TikTok, Telegram та інші, щодня обробляють мільйони повідомлень користувачів у вигляді постів, коментарів, відповідей та описів.

Станом на квітень 2025 року кількість користувачів соціальних мереж у світі досягла 5,31 мільярда, що становить 64,7% від загального населення планети[1]. Протягом останнього року до соціальних мереж приєдналося 241 мільйонів нових користувачів, що відповідає середньому темпу зростання 4,7% на рік.

У середньому користувачі щомісяця активно використовують 6,83 різних соціальних мережі[2] і щодня витрачають на них понад 2 години 20 хвилин[3], генеруючи неймовірну кількість текстового контенту.

В подальшому ці текстові дані використовуються в різних галузях досліджень[4]:

- Охорона здоров'я: аналіз повідомлень у соціальних мережах дозволяє відстежувати громадське здоров'я, виявляти поширення захворювань та оцінювати ефективність медичних кампаній.
- Політика: дослідники використовують дані соціальних мереж для аналізу громадської думки, прогнозування результатів виборів та вивчення політичної активності громадян.
- Економіка та маркетинг: соціальні мережі слугують джерелом інформації про споживчі вподобання, що дозволяє компаніям адаптувати свої стратегії та продукти до потреб ринку.
- Соціологія та поведінкові науки: аналіз поведінки користувачів у соціальних мережах допомагає вивчати соціальні взаємодії, формування спільнот та розповсюдження інформації.

Таким чином, соціальні мережі є не лише платформами для спілкування, але й потужним інструментом для збору та аналізу текстових даних у різних наукових сферах.

1.2. Проблеми та особливості обробки текстів у соціальних медіа

На відміну від традиційних текстів (наукових статей, книг чи новинних публікацій), тексти в соціальних мережах характеризуються високою динамікою, неструктурованістю та емоційністю.

Ключовими особливостями текстових даних у соціальних мережах[5] є:

- Великий обсяг: потік повідомлень є постійним і практично нескінченним. Це створює як перевагу (наявність великого масиву для навчання моделей), так і складність (потреба в ефективних інструментах обробки великих даних).
- Короткий формат повідомлень: багато платформ, особливо Twitter (X), обмежують кількість символів у повідомленні. Це призводить до стислого викладення думок, часто без чіткої структури або контексту.

- Неформальний стиль мовлення: у текстах часто використовуються розмовна лексика, сленг, скорочення, емодзі, хештеги та згадки (наприклад, *@username, #подія*), що ускладнює лінгвістичний аналіз.
- Суб'єктивність та емоційність: повідомлення часто виражають особисту думку, почуття або реакцію на події, що може супроводжуватись іронією або сарказмом.
- Тематика повідомлень є надзвичайно широкою: від політики та економіки до гумору, особистих переживань, маркетингу та культури, багато повідомлень мають міжтемний характер.
- Контекстуальна залежність: значення повідомлення часто залежить від поточних подій, попередніх постів або загального інформаційного середовища.

Усе вищезазначене свідчить про те, що соціальні мережі є надзвичайно складним джерелом даних для дослідників, які працюють у галузі обробки природної мови (NLP). Надійна класифікація таких текстів потребує підходів, що враховують їх специфіку — зокрема, здатність моделі розуміти контекст, працювати з короткими та неформальними фрагментами, а також адаптуватись до нових тем.

1.3. Методи тематичної класифікації текстових повідомлень

Тематична класифікація тексту (topic classification) — це процес автоматичного визначення теми або категорії, до якої належить певне текстове повідомлення. У контексті соціальних мереж така класифікація дозволяє структурувати величезні обсяги неструктурованого контенту, виявляти актуальні тенденції, аналізувати громадську думку та реалізовувати системи рекомендацій або модерації контенту.

Сучасні методи тематичної класифікації поділяють на кілька основних груп залежно від підходу до обробки тексту:

1. Лінгвістичні (rule-based) методи

Ці методи[6] базуються на створенні вручну прописаних правил, ключових слів, словників або регулярних виразів. Наприклад, якщо повідомлення містить слова «футбол», «матч», «гравець» — воно може бути класифіковане як спортивне.

Перевагами цього методу є простота, контрольованість, швидка реалізація для вузьких задач, недоліками – погана масштабованість, залежність від мови, низька адаптивність до змінної лексики.

2. Методи на основі класичного машинного навчання

Ці підходи передбачають векторизацію тексту (наприклад, TF-IDF, Bag of Words) і подальше застосування алгоритмів класифікації, таких як[7]:

- наївний баєсівський класифікатор (Naive Bayes);
- логістична регресія (Logistic Regression);
- дерева рішень (Decision Trees).

Моделі навчаються на розмічених даних та створюють гіперплощини або функції ймовірності, що дозволяють відносити нові тексти до певних тем.

Серед переваг можна відзначити зрозумілість моделей, швидке навчання, ефективність на малих обсягах даних, серед недоліків – недостатню здатність моделювати контекст, особливо в коротких і неформальних текстах.

3. Тематичне моделювання (Topic Modeling)

На відміну від класифікації, ці методи[8] не вимагають наявності розмічених даних, натомість намагаються виявити латентні теми в корпусі текстів. Найбільш відомим методом є LDA (Latent Dirichlet Allocation), який виявляє ймовірнісні розподіли тем у документах та слів у темах.

Його перевагами є можливість роботи з нерозміченими даними й інтерпретовані результати. Недоліками ж є те, що він не підходить для коротких текстів без модифікацій і вимагає ручного налаштування кількості тем.

4. Нейромережеві методи

Ці методи базуються на представленнях тексту у вигляді векторів із використанням моделей глибокого навчання. Серед найпоширеніших підходів[9] можна виокремити Word Embeddings + LSTM / GRU, що передбачає

представлення слів за допомогою Word2Vec, GloVe або FastText і подальшу обробку рекурентними мережами, та Transformer-моделі (наприклад, BERT, RoBERTa), що використовують контекстуальні подання, які враховують не лише сусідні слова, але й повний контекст речення.

Їх перевагами є висока точність, здатність розуміти контекст, масштабованість, проте такі моделі потребують великих обчислювальних ресурсів, є складними в навчанні та інтерпретації.

5. Zero-shot класифікація

Цей підхід[10] дозволяє класифікувати текст без попереднього навчання на конкретних тематичних класах. Модель (наприклад, zero-shot BERT або NLI-моделі) отримує текст і список потенційних тем, після чого оцінює ймовірність належності тексту до кожної теми.

Такий метод не потребує анотованих даних і є гнучким при роботі з новими або змінними тематиками, проте його точність є нижчою у порівнянні з навчуваними моделями.

Таким чином, вибір методу тематичної класифікації залежить від наявності розмічених даних, обсягу текстів, обчислювальних ресурсів та цілей дослідження.

Висновки до розділу 1

У цьому розділі було здійснено теоретичний огляд особливостей текстових повідомлень у соціальних мережах та сучасних методів їх автоматичної класифікації. Розглянуто унікальні характеристики соціального контенту, що значно ускладнюють аналіз — зокрема, короткість повідомлень, неформальний стиль, емоційність і контекстуальна залежність.

На основі цього було встановлено, що традиційні підходи обробки текстів є недостатніми для роботи з такими даними, і потребують доповнення або заміни більш гнучкими та контекстно-чутливими моделями.

У межах огляду методів класифікації було проаналізовано п’ять основних груп підходів: лінгвістичні rule-based системи, класичні ML-моделі, тематичне моделювання (LDA), нейромережеві підходи (BERT) та zero-shot класифікація.

РОЗДІЛ 2. Аналіз предметної області і постановка задачі

Перш ніж перейти до побудови та оцінювання моделей класифікації текстів, необхідно детально проаналізувати специфіку предметної області, яку становлять соціальні мережі та їхній текстовий контент. Тематична класифікація повідомлень у соціальних медіа вимагає врахування особливостей контенту, типових тематик, стилістичних характеристик текстів, а також викликів, пов’язаних із багатозначністю, неформальністю та мінливістю мови користувачів.

У цьому розділі розглядається структура типових повідомлень у соціальних мережах і виділяються основні теми, які можуть бути об'єктом класифікації. На основі цього формулюється задача автоматичної тематичної класифікації, а також визначаються вимоги до системи, яка має реалізовувати цю функціональність. Особливу увагу приділено метрикам оцінювання якості класифікації — таким, як точність (accuracy), повнота (recall), F1-міра та інші — з обґрунтуванням їх застосування в контексті тематичного аналізу повідомлень.

Матеріал цього розділу формує методологічну основу для подальшої реалізації та тестування моделей класифікації в практичній частині роботи.

2.1. Аналіз типів повідомлень у соціальних мережах за темами

Соціальні мережі є середовищем активної взаємодії користувачів, які щодня публікують мільйони повідомлень на різноманітні теми. Для побудови ефективної системи тематичної класифікації таких повідомлень важливо визначити релевантні категорії, що відображають найпоширеніші напрями онлайн-комунікації.

У цьому дослідженні було виокремлено 20 основних тематичних класів повідомлень: технології, політика, спорт, кіно, ігри, мистецтво, музика, здоров’я, їжа, бізнес, фінанси, освіта, наука, подорожі, стосунки, батьківство, екологія, кар’єра, домашні тварини, фітнес.

Такий вибір базується на поєднанні емпіричного аналізу соціального контенту, попередніх досліджень, а також глобальних цифрових звітів.

2.1.1. Обґрунтування вибору тематичних категорій

Згідно з аналітичним звітом DataReportal[11], 2024, найпоширенішими інтересами користувачів у соціальних медіа є розваги (кіно, музика, ігри), спорт, здоров'я, новини (зокрема, політика), технології, мандрівки та кулінарія. Таким чином, значна частина виокремлених тем прямо відображає реальні запити та активність користувачів.

Дослідження Srijith і Hepple (2017)[12] показують, що ефективна класифікація повідомлень у Twitter та інших соцмережах може спиратись на теми на кшталт: «технології», «політика», «бізнес», «здоров'я», «спорт», «розваги», «наука» — що корелює з обраним набором у цій роботі.

Також виокремлені теми охоплюють як особисту сферу життя користувача (стосунки, батьківство, кар'єра, фітнес, домашні тварини), так і глобальні сфери знань і суспільного інтересу (політика, фінанси, наука, екологія). Це забезпечує комплексний підхід до аналізу текстового контенту в соціальних мережах.

Теми були підібрані так, щоб вони були максимально відмежовані одна від одної в лексичному, семантичному та стилістичному вимірах, що дозволяє моделі краще диференціювати повідомлення.

Таким чином, вибір саме цих 20 категорій дозволяє сформувати тематичну структуру, яка є водночас достатньо репрезентативною та актуальною для широкого кола реальних повідомлень у соціальних мережах.

2.2. Постановка задачі тематичної класифікації повідомлень

Автоматична тематична класифікація повідомлень у соціальних мережах полягає у визначенні, до якої з наперед заданих категорій належить те чи інше текстове повідомлення. Ця задача є різновидом проблеми класифікації в галузі обробки природної мови (NLP) та має особливе значення в умовах сучасного

інформаційного середовища, де щодня генерується великий обсяг неструктурованого контенту.

Зважаючи на специфіку текстів у соціальних мережах (короткість, неформальність, контекстуальність), ця задача набуває додаткової складності. Повідомлення часто не містять явних індикаторів теми, а самі теми можуть бути змішаними. У зв'язку з цим класифікація текстів із соціальних мереж потребує уваги до попередньої обробки даних, вибору архітектури моделі та метрик оцінки якості.

2.3. Вимоги до систем класифікації

Розробка системи автоматичної тематичної класифікації повідомлень у соціальних мережах потребує чіткого визначення вимог, які повинна задовольняти така система. Вимоги визначаються як технічними, так і лінгвістичними особливостями текстів, а також очікуваними сценаріями застосування.

Умовно всі вимоги можна поділити на функціональні та нефункціональні.

2.3.1. Функціональні вимоги

1. Наявність попередньої обробки тексту:
   - система повинна підтримувати очищення текстів від шумових символів (емодзі, HTML-тегів, згадок @user, хештегів);
   - передбачена нормалізація, токенізація, лемматизація або стемінг.
2. Наявність векторизації тексту:
   - необхідна трансформація текстових даних у числові представлення (наприклад, TF-IDF, Word2Vec, BERT embeddings).
3. Оцінка якості класифікації:
   - система повинна повертати метрики точності, повноти, F1-міри, а також візуалізації результатів (наприклад, confusion matrix).

2.3.2. Нефункціональні вимоги

1. Масштабованість: система має підтримувати обробку великого обсягу повідомлень, характерного для реального потоку в соцмережах.
2. Продуктивність: час класифікації одного повідомлення не повинен перевищувати допустимі значення для задач реального часу (особливо для zero-shot inference).
3. Інтерпретованість: результати класифікації мають бути зрозумілими для кінцевого користувача або аналітика (наприклад, через показ найбільш релевантних слів/тем для кожного класу).
4. Стійкість до шумів: модель повинна зберігати стабільність роботи при наявності орфографічних помилок, сленгу та інших нетипових мовних елементів.

Виконання вищезазначених вимог є важливим для створення надійної системи тематичної класифікації повідомлень у соціальних мережах, здатної працювати в умовах реального динамічного середовища, змінних мовних конструкцій та високого обсягу вхідних даних.

2.4. Обґрунтування вибору метрик якості класифікації

Для оцінювання ефективності тематичної класифікації повідомлень у соціальних мережах необхідно використовувати формалізовані метрики якості, що дозволяють об'єктивно порівнювати різні моделі та визначати їх придатність до реальних умов. Через особливості задачі важливо застосовувати комплексний підхід до оцінювання, що виходить за межі лише однієї метрики.

У межах цієї роботи було використано такі метрики:

1. Accuracy (Точність класифікації)

$$Accuracy = \frac{\text{Кількість правильних передбачень}}{\text{Загальна кількість передбачень}}$$

Переваги: проста у розрахунку, дає загальне уявлення про точність моделі.

Недоліки: не підходить для задач із незбалансованими класами, коли один клас значно переважає інші. Наприклад, модель може показувати високу точність, просто класифікуючи більшість прикладів як найбільш популярну тему.

2. Precision (Точність)

Показує, яка частка передбачених позитивних випадків дійсно належить до цільового класу. Корисна, коли важливо мінімізувати хибнопозитивні спрацьовування — наприклад, у задачах модерації або фільтрації.

Для кожного класу C_i точність обчислюється так:

$$Precision_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + FP_{C_i}}$$

де:

- TP_{C_i} — кількість правильно передбачених прикладів класу C_i,
- FP_{C_i} — кількість прикладів, які були неправильно віднесені до C_i.

Далі загальна точність числюється як середнє арифметичне точності по всіх класах:

$$Precision_{macro} = \frac{1}{K}\sum_{i=1}^{K} Precision_{C_i}$$

де:

K — кількість класів.

3. Recall (Повнота)

Визначає, яку частку справжніх позитивних прикладів було правильно виявлено. Критична, коли важливо не пропустити релевантні приклади — наприклад, у виявленні важливих новинних тем.

Для кожного класу C_i повнота обчислюється так:

$$Recall_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + FN_{C_i}}$$

де:

- TP_{C_i} — кількість правильно передбачених прикладів класу C_i,
- FN_{C_i} — кількість прикладів, які були помилково не віднесені до C_i.

Загальна повнота обчислюється аналогічно до загальної точності (precision).

4. F1-score (Гармонійне середнє між precision і recall)

Найбільш збалансована метрика для оцінювання якості класифікації в умовах нерівномірного розподілу класів. Рекомендована як основна метрика у багатокласовій класифікації коротких соціальних повідомлень.

Формула для обчислення:

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

5. Confusion Matrix (Матриця помилок)

Надає детальну інформацію про те, які класи модель плутає між собою. Це дозволяє виявити слабкі місця, покращити розмітку даних або баланс класів та інтерпретувати результати у прикладному контексті.

6. Macro- та Micro- усереднення

У багатокласовій класифікації важливо враховувати спосіб усереднення:

- Macro-average — розраховує метрику незалежно для кожного класу і потім усереднює. Підходить, коли всі класи однаково важливі.
- Micro-average — агрегує усі true positives, false negatives та false positives і рахує метрику на основі сум. Краще для задач із незбалансованими класами.

Отже, у цьому дослідженні для оцінювання якості моделей тематичної класифікації використовуються такі основні метрики:

- Accuracy — як базовий орієнтир;
- Precision, Recall, F1-score — як ключові метрики для детального аналізу;
- F1-macro та F1-micro — для врахування впливу класового дисбалансу;
- Confusion matrix — для візуального аналізу помилок моделі.

Такий підхід дозволяє не лише отримати числову оцінку якості, а й зробити глибший аналіз результатів і поведінки моделей у контексті реальних даних соціальних мереж.

2.5. Обґрунтування вибору метрик якості кластеризації

Оцінювання якості кластеризації текстових повідомлень у соціальних мережах є нетривіальним завданням, особливо в умовах відсутності попередньої ручної розмітки або еталонних міток (ground truth). У таких випадках стандартні метрики точності не можуть бути безпосередньо застосовані, що зумовлює потребу в альтернативних способах оцінки.

У межах даного дослідження для оцінки результатів кластеризації використовувався візуально-інтерпретативний підхід на основі емпіричного аналізу. Основними інструментами стали:

- UMAP (Uniform Manifold Approximation and Projection) — метод зниження розмірності, який дозволяє зберегти локальні та глобальні структури високовимірних ембедінгів у 2D або 3D-просторі;
- Візуалізація кластерів — графічне відображення результатів кластеризації в зниженому просторі, що дозволяє аналізувати щільність, відокремленість і змістовність кластерів.

Цей підхід дозволив:

- Визначити якісну сегментацію повідомлень — тобто, наскільки тісно згруповані об'єкти одного кластера і наскільки чітко відмежовані вони від інших кластерів;
- Провести інтерпретацію кластерів на основі частотного аналізу слів і типових повідомлень;
- Зіставити отримані кластери з наявними категоріями або тематичними очікуваннями (наприклад, батьківство, харчування, музика тощо).

Таким чином, оцінювання кластеризації здійснювалося на основі візуальної інспекції розподілу кластерів у UMAP-просторі та семантичного аналізу їх вмісту. Такий підхід, попри відсутність формальних числових метрик, дозволив забезпечити практично релевантну і змістовну оцінку якості кластеризації, що є прийнятним у задачах тематичного групування неструктурованих текстів без попереднього маркування.

Висновки до розділу 2

У другому розділі було здійснено системний аналіз предметної області, яка охоплює тематичну класифікацію текстових повідомлень у соціальних мережах. Зважаючи на специфіку соціального контенту — його неструктурованість, короткість, емоційність, контекстуальність та тематику — було обґрунтовано необхідність застосування адаптивних методів обробки природної мови, зокрема на основі сучасних алгоритмів машинного навчання та нейромереж.

Було здійснено такі ключові кроки:

1. Проаналізовано типи повідомлень, характерних для соціальних мереж, та сформовано 20 тематичних категорій, що охоплюють як особисту, так і суспільно важливу тематику. Їх вибір було обґрунтовано на основі емпіричного аналізу, глобальних статистичних звітів та актуальних наукових досліджень.
2. Сформульовано задачу тематичної класифікації.
3. Визначено функціональні та нефункціональні вимоги до системи тематичної класифікації. Особлива увага приділена таким аспектам, як попередня обробка тексту, масштабованість, гнучкість, стійкість до шуму та інтерпретованість результатів.
4. Обґрунтовано вибір метрик для оцінки якості класифікації, серед яких основними є: accuracy, precision, recall, F1-score (у варіантах macro та micro), а також confusion matrix для візуального аналізу помилок.
5. У контексті кластеризації повідомлень було обрано візуальні методи оцінки якості, зокрема UMAP для зниження розмірності та подальшої інтерпретації кластерів. Акцент було зроблено на графічному аналізі щільності, відокремленості та змістовної узгодженості кластерів, а також на семантичному аналізі ключових слів і прикладів повідомлень у кожному кластері. Такий підхід дозволив ефективно оцінити якість кластеризації без потреби в еталонній розмітці.

Таким чином, розділ 2 закладає теоретичну та методологічну основу для побудови, навчання та порівняльного аналізу моделей тематичної класифікації, що буде реалізовано у наступних етапах дослідження.

РОЗДІЛ 3. Реалізація підходів на основі класичних алгоритмів машинного навчання

Класичні алгоритми машинного навчання залишаються популярним підходом до задач тематичної класифікації текстів завдяки своїй простоті, швидкості та ефективності при роботі з невеликими або помірними обсягами даних. У цьому розділі розглядається реалізація трьох базових моделей класифікації текстових повідомлень із соціальних мереж на основі TF-IDF-векторизації у поєднанні з різними алгоритмами: наївним баєсівським класифікатором, логістичною регресією та методом опорних векторів (SVM).

Основна мета цього розділу — дослідити, якої точності можна досягти за допомогою класичних підходів у багатокласовій класифікації коротких повідомлень із соціальних мереж, а також виявити обмеження, з якими стикаються ці моделі при обробці неформального та контекстно-залежного контенту.

3.1. Опис та структура датасету

Для навчання та тестування моделей тематичної класифікації було використано власноруч сформований датасет, зібраний із платформи Reddit[13]. Цей ресурс є цінним джерелом текстових даних, оскільки кожен тематичний підрозділ (subreddit) структуровано навколо певної теми, що значно спрощує процес тематичного маркування.

Усього датасет містить 20 000 текстових повідомлень. Кожен запис — це коротке повідомлення (один пост), яке представлено в один рядок. Більшість повідомлень — це речення або абзаци до 2–3 рядків, тобто типові короткі тексти соціальних мереж.

Категорії були відібрані на основі аналізу актуальних тем у соціальних мережах (описано в розділі 2.1.1.) і охоплюють широкий спектр інтересів користувачів: технології, політика, спорт, кіно, ігри, мистецтво, музика, здоров’я, їжа, бізнес, фінанси, освіта, наука, подорожі, стосунки, батьківство, екологія, кар’єра, домашні тварини, фітнес.

Кількість прикладів у кожному класі є однаковою — по 1000 повідомлень, що дозволяє говорити про збалансований датасет. Для візуалізації цього розподілу було побудовано графік (рис. 3.1), що ілюструє його рівномірність.

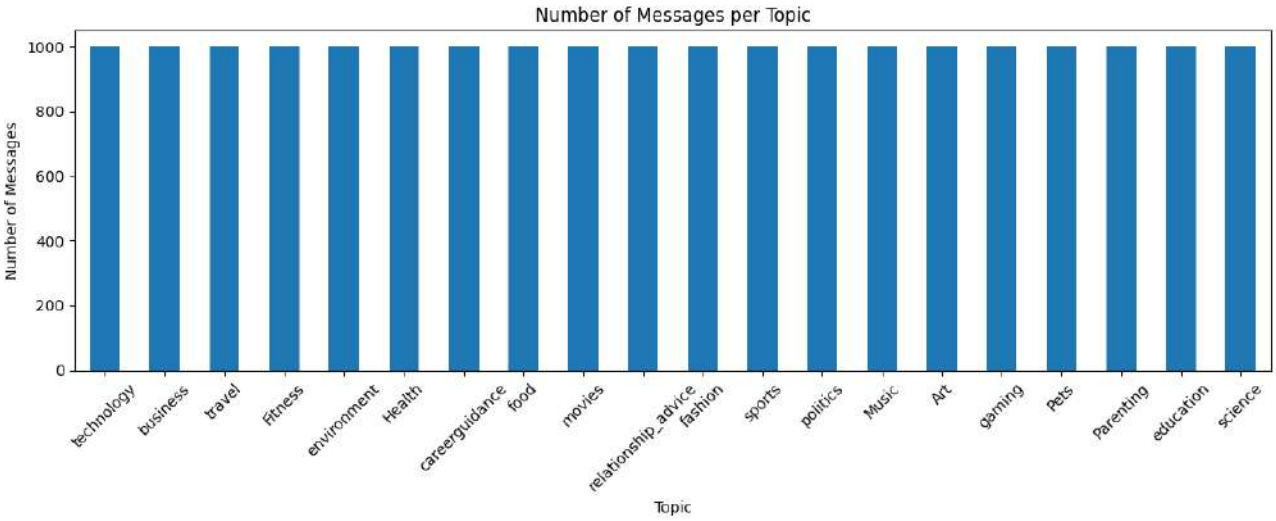

Рисунок 3.1. Кількість повідомлень для кожної категорії датасету

Перед подачею текстів у моделі було проведено стандартні етапи попередньої обробки (preprocessing), що включали:

1. Очищення тексту: видалення HTML-символів, згадок користувачів (@user), емодзі, хештегів (#), зайвих пробілів та пунктуації.
2. Зниження регістру: усі тексти було приведено до нижнього регістру для уніфікації лексем.
3. Токенізація та лемматизація: розбиття тексту на окремі слова (токени) та приведення слів до їхніх базових форм (лем). Лемматизація допомогла зменшити розмір словника і збільшити узагальнюваність моделі.
4. Видалення стоп-слів: було усунено найпоширеніші службові слова (наприклад, "and", "or", "also"), які не несуть суттєвого змісту для класифікації.

Приклад результатів очищення і токенізації поданий на рисунку 3.2.

	text	label	tokens
10549	Is this jacket too short or just right?	fashion	[jacket, short, right]
8938	Double features in the cinema? I haven't been ...	movies	[double, feature, cinema, able, afford, go, ci...
19310	For Li-ion battery SOC prediction using machin...	science	[li, ion, battery, soc, prediction, use, machi...
4943	EPA plans to eliminate scientific research tea...	environment	[epa, plan, eliminate, scientific, research, t...
6604	How do I move states when all my professional ...	careerguidance	[move, state, professional, connection, live, ...

Рисунок 3.2. Результати очищення і токенізації повідомлень

Таким чином, підготовлений датасет є якісною базою для навчання моделей класифікації, з чітко визначеними темами, збалансованим розподілом і повноцінною попередньою обробкою тексту. У наступних підрозділах описується реалізація та результати застосування класичних алгоритмів машинного навчання до цих даних.

3.2. Векторизація тексту

Для подання текстових повідомлень у вигляді, придатному для обробки класичними алгоритмами машинного навчання, необхідно здійснити їх числове представлення. У цьому дослідженні для цього було застосовано метод TF-IDF[14] (Term Frequency-Inverse Document Frequency) — один із найпоширеніших способів векторизації тексту в задачах обробки природної мови.

Метод TF-IDF дозволяє обчислити вагу кожного слова в документі з урахуванням того, наскільки часто воно зустрічається в цьому документі (TF) і наскільки рідкісним воно є в усьому корпусі (IDF).

Таким чином, цей підхід підсилює вплив термінів, які є характерними для конкретного повідомлення, і зменшує вагу загальних слів, що часто трапляються в багатьох документах. Це дозволяє моделі зосередитись на дійсно важливих словах при класифікації.

Для реалізації TF-IDF-векторизації було використано інструмент TfidfVectorizer з бібліотеки scikit-learn.

Після векторизації було сформовано розріджену матрицю розміром:

- X_train_tfidf.shape = (16000, 10000);
- X_test_tfidf.shape = (4000, 10000).

Це означає, що кожне повідомлення було перетворено на вектор з 10 000 ознак, де кожна ознака відповідає певному слову з обраного словника.

Нижче наведено приклад одного з повідомлень у векторизованому вигляді (рис. 3.3).

```
Original text:
apple approve spotify app update external payment follow update spotify freely advertise price user subscribe outside app

Most important words:
spotify: 0.4683
update: 0.3523
app: 0.3497
advertise: 0.2776
subscribe: 0.2682
freely: 0.2632
external: 0.2549
payment: 0.2235
approve: 0.2196
user: 0.2026
apple: 0.1942
price: 0.1798
outside: 0.1552
follow: 0.1331
```

Рисунок 3.3. Приклад векторизації повідомлення за допомогою TF-IDF

Як видно, модель призначила найвищі ваги словам "spotify", "update", "app", які є ключовими для розуміння змісту повідомлення. Це демонструє здатність TF-IDF фокусуватись на найбільш релевантних термінах та відкидати малозначущі, що підвищує якість подальшої класифікації.

Таким чином, TF-IDF векторизація забезпечила оптимальне, розріджене та інформативне представлення текстів, що є придатною основою для подальшого навчання моделей класифікації на основі класичних алгоритмів машинного навчання.

3.3. Модель 1: TF-IDF + Naive Bayes

У першому експерименті для класифікації повідомлень у соціальних мережах було застосовано класичний підхід: поєднання TF-IDF-векторизації тексту та моделі Multinomial Naive Bayes[15].

Наївний баєсівський класифікатор є простим і швидким алгоритмом, який добре працює з розрідженими матрицями ознак, що утворюються після TF-IDF-векторизації. Він часто демонструє непогану якість при мінімальних обчислювальних витратах, що робить його ефективним для задач з великою кількістю класів. У цьому дослідженні модель використовується як базовий орієнтир для подальшого порівняння з іншими підходами.

Для векторизації тексту було використано TfidfVectorizer з такими параметрами:

- max_features=10000 — обмеження словника 10 000 найбільш інформативними словами;
- ngram_range=(1, 3) — врахування уніграм, біграм і триграм;
- min_df=2 — виключення дуже рідкісних слів;
- stop_words='english' — видалення англійських стоп-слів.

Модель MultinomialNB була налаштована з параметром alpha=0.3, який забезпечив кращий баланс між згладжуванням і чутливістю моделі до частоти слів.

Нижче наведені метрики для кожного з класів (рис. 3.4).

Accuracy: 0.81925
Classification Report:

	precision	recall	f1-score	support
Art	0.97	0.89	0.93	194
Fitness	0.99	0.97	0.98	203
Health	0.76	0.68	0.72	199
Music	0.89	0.78	0.83	209
Parenting	0.78	0.81	0.79	200
Pets	0.88	0.94	0.91	202
business	0.61	0.53	0.57	179
careerguidance	0.76	0.96	0.85	211
education	0.73	0.83	0.78	194
environment	0.80	0.73	0.76	214
fashion	0.92	0.88	0.90	194
food	0.95	0.98	0.96	202
gaming	0.85	0.82	0.84	193
movies	0.84	0.91	0.88	208
politics	0.72	0.74	0.73	204
relationship_advice	0.83	0.95	0.89	204
science	0.71	0.66	0.68	188
sports	0.88	0.74	0.80	203
technology	0.58	0.61	0.60	173
travel	0.90	0.90	0.90	226
accuracy			0.82	4000
macro avg	0.82	0.82	0.81	4000
weighted avg	0.82	0.82	0.82	4000

Рисунок 3.4. Результати класифікації за допомогою Naive Bayes

Після аналізу результатів класифікації було виявлено такі сильні сторони:

- Високі результати для чітко окреслених тем (fitness, food, art, pets).
- Врахування біграм і триграм (ngram_range=(1, 3)) покращило точність у багатьох класах.
- Простота моделі дозволяє її швидко тренувати і застосовувати навіть на великих обсягах даних.

Виявленими слабкими сторонами були:

- Слабша ефективність у класах зі змішаною тематикою або абстрактною лексикою (business, health, science).
- У класах на кшталт technology та business — нижчі значення precision і recall, що вказує на змішування з подібними темами.

Змішування подібних тем також підтверджує побудована Confusion matrix (рис. 3.5).

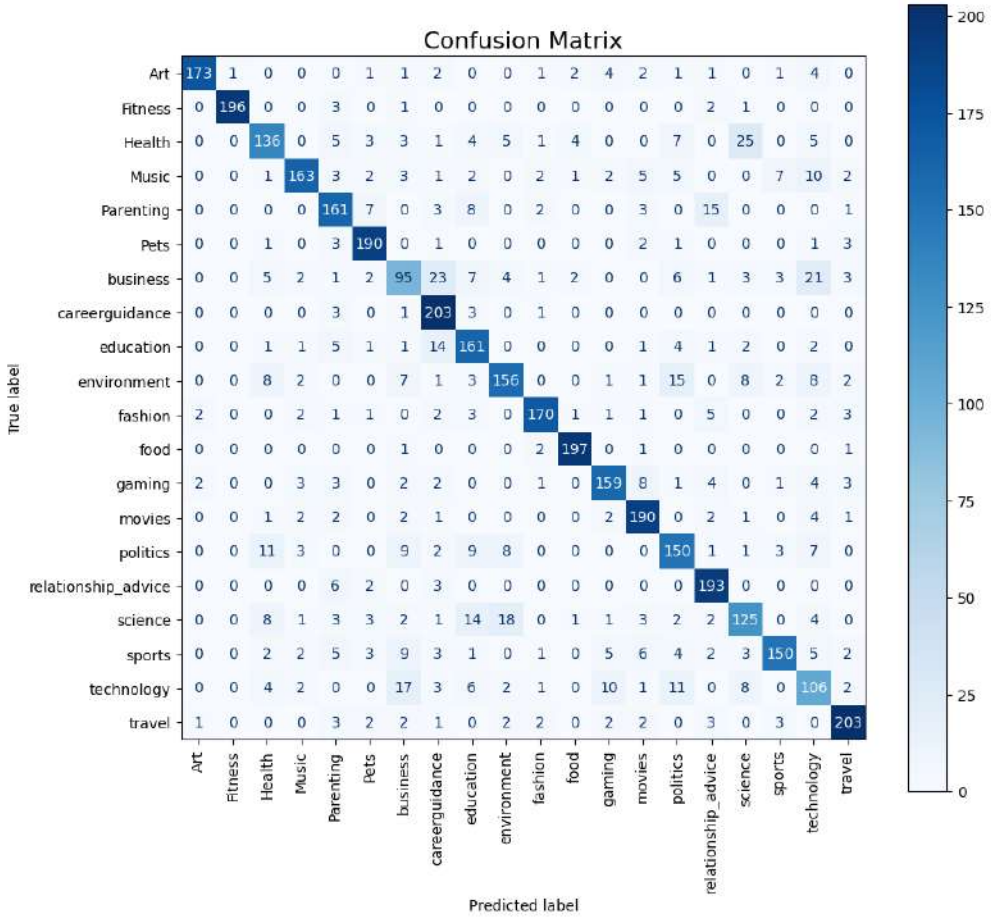

Рисунок 3.5. Confusion matrix за результатами класифікації з Naive Bayes

Як бачимо, модель плутає такі класи, як health і science (обидва можуть містити медичні терміни, назви частин тіла), business і career та business і technology, які також містять спільні терміни.

Таким чином, модель TF-IDF + Naive Bayes демонструє гідний рівень якості для базового підходу, проте її обмеження у роботі з контекстом та семантикою тексту вказують на доцільність переходу до нейромережевих архітектур, які краще адаптовані до обробки неструктурованих і неформальних текстів із соціальних мереж.

3.4. Модель 2: TF-IDF + Logistic Regression

Логістична регресія[16] — це лінійний алгоритм класифікації, який обчислює ймовірність належності об'єкта до певного класу. У багатокласовій класифікації модель застосовує підхід "один проти всіх" (One-vs-Rest), тобто навчає окремий класифікатор для кожного класу, який визначає, чи належить зразок до цього класу чи ні.

На основі лінійної комбінації ознак модель оцінює ймовірності для всіх класів і обирає той, що має максимальне значення. Завдяки цьому підхід добре працює з високовимірними даними, зокрема з TF-IDF-векторами тексту.

Для побудови моделі логістичної регресії використовувались наступні параметри:

- TfidfVectorizer з параметрами:
  - max_features=15000 — словник із 15 000 найбільш інформативних слів;
  - min_df=3 — ігнорування дуже рідкісних слів;
  - stop_words='english' — фільтрація англійських стоп-слів.
- Модель логістичної регресії:
  - C=5 — зменшена сила регуляризації (вища гнучкість моделі);
  - penalty='l2' — класична L2-регуляризація;

- solver='liblinear' — оптимізатор, який підходить для невеликих і середніх наборів даних;
- max_iter=1000 — збільшене число ітерацій для гарантії збіжності.

Рисунок 3.6 подає метрики класифікації за основними показниками.

	precision	recall	f1-score	support
Art	0.96	0.93	0.94	194
Fitness	1.00	0.99	0.99	203
Health	0.78	0.72	0.75	199
Music	0.86	0.90	0.88	209
Parenting	0.87	0.84	0.86	200
Pets	0.96	0.96	0.96	202
business	0.62	0.65	0.64	179
careerguidance	0.91	0.88	0.90	211
education	0.84	0.82	0.83	194
environment	0.76	0.76	0.76	214
fashion	0.92	0.91	0.91	194
food	0.98	0.97	0.97	202
gaming	0.88	0.86	0.87	193
movies	0.94	0.88	0.91	208
politics	0.76	0.80	0.78	204
relationship_advice	0.95	0.93	0.94	204
science	0.67	0.68	0.68	188
sports	0.77	0.86	0.81	203
technology	0.58	0.62	0.60	173
travel	0.89	0.89	0.89	226
accuracy			0.85	4000
macro avg	0.85	0.84	0.84	4000
weighted avg	0.85	0.85	0.85	4000

Рисунок 3.6. Результати класифікації за допомогою Logistic Regression

За результатами аналізу виявлено такі позитивні аспекти:

- Модель досягла вищої загальної точності (85%), ніж базовий класифікатор наївного Баєса.
- Високі значення F1-міри у більшості класів, зокрема Fitness, Food, Pets, Art — свідчать про точність і стабільність моделі.
- Логістична регресія краще впоралась з класами з перетинною або подібною лексикою, зокрема movies, music, parenting.

Обмеження:

- Класи technology, business та science залишаються найскладнішими для моделі (рис. 3.7): мають нижчі значення precision і recall, що вказує на можливі тематичні перетини або брак виразної лексики.
- Хоч модель і підтримує багатокласову класифікацію, вона не моделює послідовність слів, тому не враховує контекст у повному розумінні (на відміну від нейромережевих підходів).

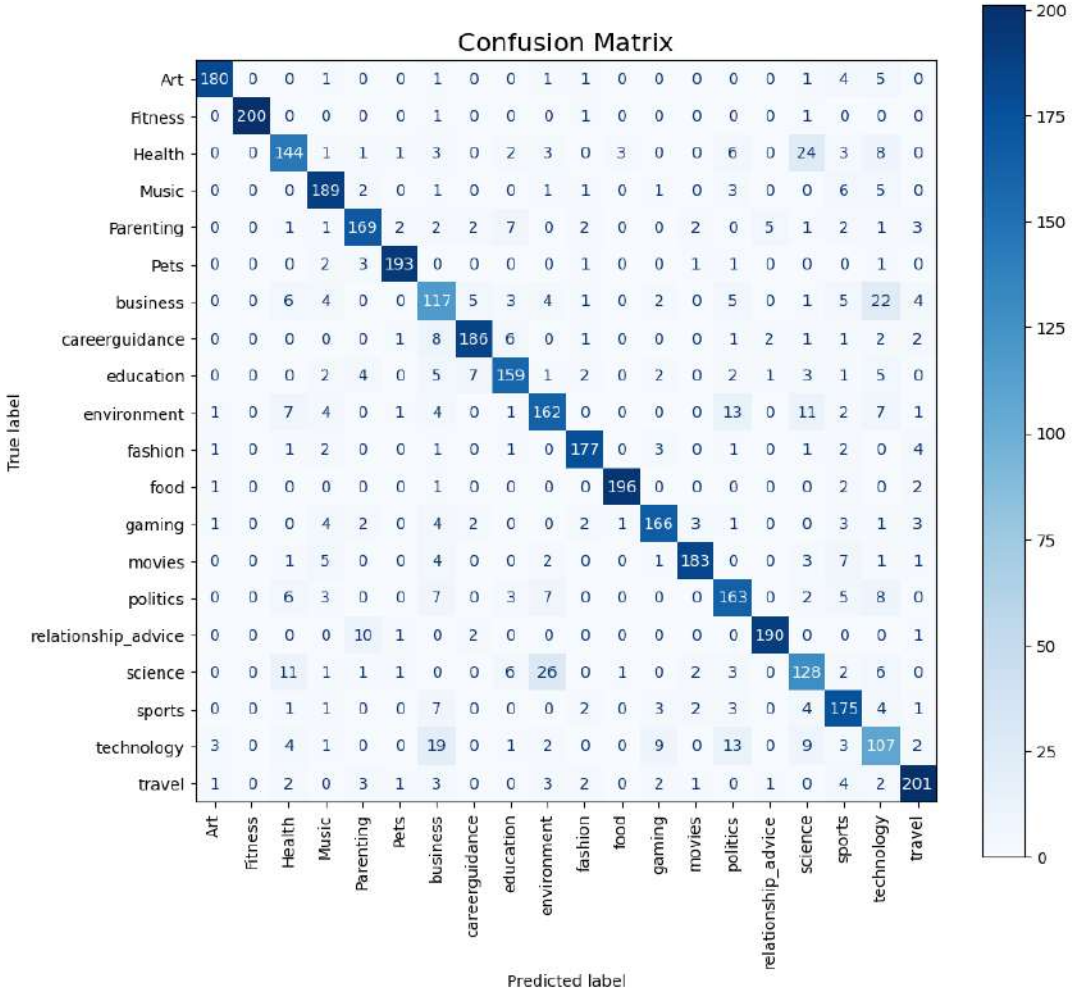

Рисунок 3.7. Confusion matrix за результатами класифікації з Logistic Regression

Отже, логістична регресія у поєднанні з TF-IDF-векторизацією забезпечила високий рівень класифікації текстів у соціальних мережах. Завдяки гнучкому налаштуванню та використанню біграм модель змогла поліпшити точність у ряді тем. Утім, обмеження цього підходу у розумінні глибшого контексту також свідчать про доцільність переходу до нейромережевих архітектур, таких як LSTM або BERT, у наступних етапах дослідження.

3.5. Модель 3: TF-IDF + SVM

Метод опорних векторів[17] (Support Vector Machine, SVM) є одним із найпотужніших інструментів для задач класифікації, особливо при роботі з високовимірними ознаками. SVM створює гіперплощину, яка максимально розділяє класи, і добре справляється навіть у ситуаціях, коли кількість ознак значно перевищує кількість зразків.

У задачах тематичної класифікації повідомлень із соціальних мереж SVM демонструє високу точність і стабільність за умови правильного налаштування гіперпараметрів.

У реалізації було використано LinearSVC — оптимізовану версію SVM для лінійної класифікації, яка значно швидша за класичний SVC(kernel='linear'), особливо при великій кількості ознак, як у випадку TF-IDF.

Для побудови моделі було застосовано такі налаштування:

- Векторизація (TF-IDF):
  - max_features=18000 — використано 18 000 найчастотніших ознак;
  - ngram_range=(1,1) — лише уніграми (окремі слова);
  - min_df=2 — фільтрація дуже рідкісних слів;
  - sublinear_tf=True — логарифмічне масштабування частоти термінів;
  - stop_words='english' — видалення англійських стоп-слів.
- Модель LinearSVC:
  - C=1.0 — стандартна сила регуляризації;
  - loss='squared_hinge' — квадратична функція втрат;
  - max_iter=2000 — збільшена кількість ітерацій.

Результати класифікації наведено на рисунку 3.8.

	precision	recall	f1-score	support
Art	0.96	0.92	0.94	194
Fitness	1.00	0.99	1.00	203
Health	0.78	0.75	0.77	199
Music	0.87	0.90	0.88	209
Parenting	0.87	0.87	0.87	200
Pets	0.96	0.97	0.97	202
business	0.62	0.65	0.63	179
careerguidance	0.89	0.90	0.89	211
education	0.83	0.84	0.83	194
environment	0.76	0.74	0.75	214
fashion	0.92	0.92	0.92	194
food	0.98	0.99	0.98	202
gaming	0.88	0.87	0.88	193
movies	0.95	0.92	0.93	208
politics	0.75	0.79	0.77	204
relationship_advice	0.96	0.95	0.96	204
science	0.69	0.66	0.67	188
sports	0.84	0.84	0.84	203
technology	0.57	0.59	0.58	173
travel	0.91	0.90	0.91	226
accuracy			0.85	4000
macro avg	0.85	0.85	0.85	4000
weighted avg	0.85	0.85	0.85	4000

Рисунок 3.8. Результати класифікації за допомогою SVM

Результати класифікації показали такі переваги цього підходу:

- Найвища точність серед усіх протестованих класичних моделей (вище за Naive Bayes і Logistic Regression).
- Високі F1-значення для більшості тем (Fitness, Food, Pets, Movies, Relationship Advice).
- Добре справляється з великою кількістю ознак без втрати продуктивності.

І виявили такі обмеження:

- Класи з широким або абстрактним значенням (technology, business, science) все ще мають нижчу якість — модель має труднощі з розрізненням схожих тем (рис. 3.9).
- Використання лише уніграмів (ngram_range=(1,1)) обмежує врахування контексту, проте розширення до біграм або триграм лише погіршує результат.

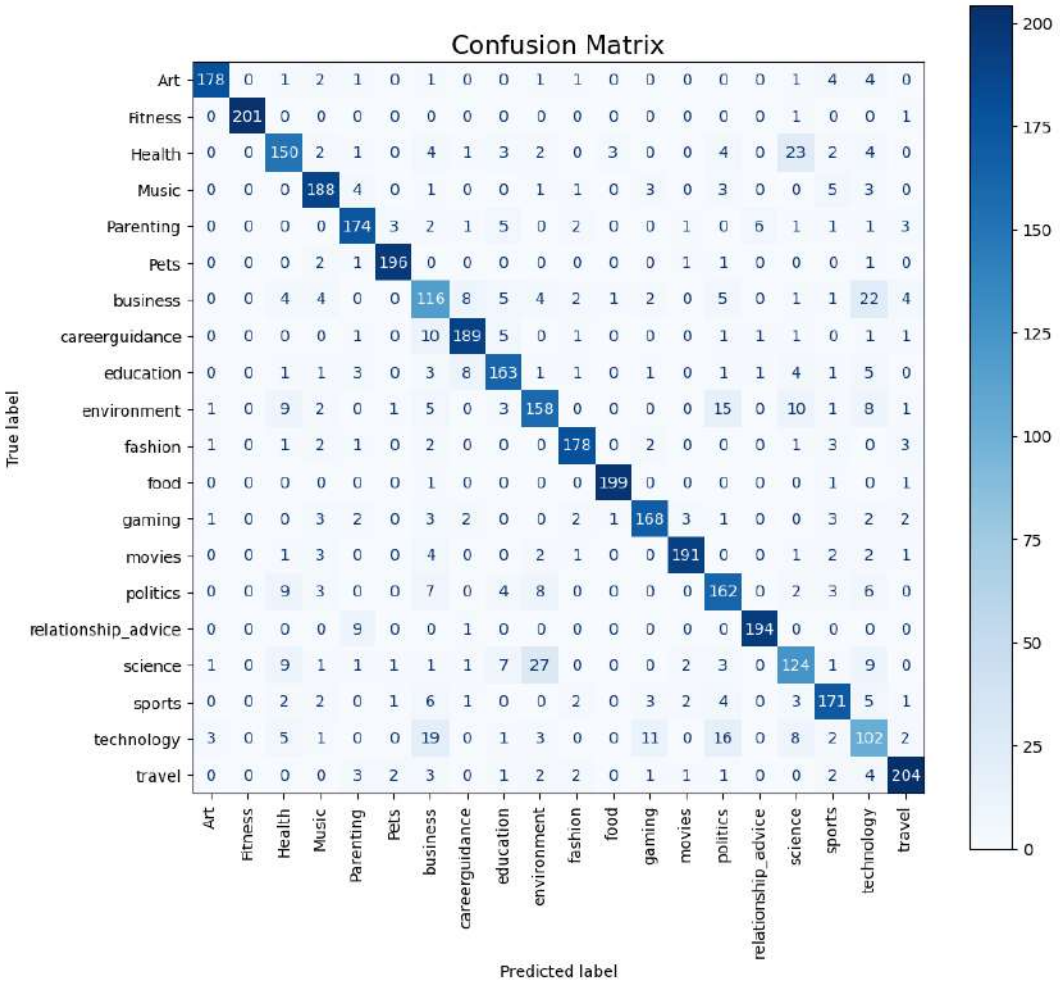

Рисунок 3.9. Confusion matrix за результатами класифікації з SVM

Отже, модель TF-IDF + SVM (LinearSVC) продемонструвала найвищу якість класифікації серед класичних підходів у цьому дослідженні. Її результати підтверджують ефективність SVM у задачах текстової класифікації з великою кількістю класів. Проте, незважаючи на високу точність, модель не враховує глибокий контекст та семантичну структуру мови, що мотивує до подальшого використання нейромережевих моделей для досягнення ще кращої якості в задачах тематичної класифікації текстів у соціальних мережах.

Проміжні висновки

У цьому розділі було реалізовано та порівняно три класичні підходи до тематичної класифікації текстових повідомлень у соціальних мережах: TF-IDF у

поєднанні з Multinomial Naive Bayes, Logistic Regression та Support Vector Machine (SVM). Усі моделі демонстрували гідний базовий рівень точності:

- Naive Bayes: accuracy = 0.819, F1 macro ≈ 0.81
- Logistic Regression: accuracy = 0.85, F1 macro ≈ 0.84
- SVM: accuracy = 0.8515, F1 macro ≈ 0.85

Однак, детальний аналіз результатів показав ряд системних обмежень класичних алгоритмів:

1. Ігнорування контексту: моделі не здатні враховувати порядок слів або їх взаємозв’язки, що знижує точність у темах з багатозначною або абстрактною лексикою (science, business, technology).
2. Слабка робота з синонімами: без контекстного подання слова як "job", "career" чи "work" сприймаються як різні, що ускладнює класифікацію.
3. Чутливість до формулювань: невеликі зміни в тексті (наприклад, заміна "like" на "enjoy") можуть призводити до зниження точності, оскільки модель орієнтується на конкретні терміни, а не їхнє значення.

Ці обмеження вказують на необхідність переходу до нейромережевих підходів, які здатні працювати з контекстом, моделювати складні мовні залежності та краще адаптуватися до варіативності текстів. Їх реалізація та оцінка будуть розглянуті в наступному розділі.

РОЗДІЛ 4. Реалізація підходів на основі сучасних алгоритмів

У цьому розділі розглянуто реалізацію сучасних підходів до тематичної класифікації текстів на основі тематичного моделювання та трансформерних моделей. На відміну від класичних алгоритмів машинного навчання, ці методи дозволяють враховувати контекст, семантику та послідовність слів, що особливо важливо для аналізу неструктурованих і неформальних текстів із соціальних мереж.

У межах розділу реалізовано три підходи: тематичне моделювання з використанням LDA (Latent Dirichlet Allocation), кластеризацію BERT-ембедінгів повідомлень за допомогою K-Means та HDBSCAN, а також zero-shot класифікацію без потреби у навчанні на конкретних тематичних даних. Кожен із методів оцінюється за якістю результатів, здатністю моделювати контекст і придатністю до різних типів задач.

4.1. Тематичне моделювання за допомогою LDA

Тематичне моделювання — це підхід до виявлення прихованих семантичних структур у великих масивах текстових даних. Одним з найпоширеніших методів є Latent Dirichlet Allocation[18] (LDA), який дозволяє автоматично виявити теми в колекції документів на основі розподілу слів у текстах.

Для застосування LDA-моделі було проведено попередню обробку текстів: лематизація, фільтрація стоп-слів, видалення чисел і пунктуації, залишено лише значущі слова. Після побудови словника та корпусу текстів було здійснено навчання моделі з параметром num_topics=20, що дозволяє виділити 20 тематичних напрямів. Також було використано налаштування для зменшення кількості тем у кожному документі (alpha=0.01) та кількості слів у кожній темі (eta=0.01), що підвищило специфічність результатів.

Значення кількості тем було визначено експериментально на основі якості розподілу тем і зрозумілості топ-слів (coherence score). Для кожної теми модель

сформувала список найбільш релевантних слів із найбільшими вагами. Наприклад:

- Тема 0 пов’язана з соціальною поведінкою: people, let, treat, behavior, stop.
- Тема 3 (рис. 4.1) охоплює політичну тематику: trump, ai, win, vote, power.
- Тема 4 стосується медицини та здоров’я: medical, routine, pain, nutrition.
- Тема 10 містить елементи попкультури: movie, music, film, dress.
- Тема 11 явно пов’язана з домашніми тваринами: dog, cat, food, house.
- Тема 13 чітко описує геймерську тематику: game, play, video, console.
- Тема 18 поєднує фітнес і лайфстайл: fitness, girl, song, hair, oil.

Інші теми також мають чітку тематичну спрямованість: освіта, емоції, мандрівки, самообслуговування, робота, мистецтво тощо.

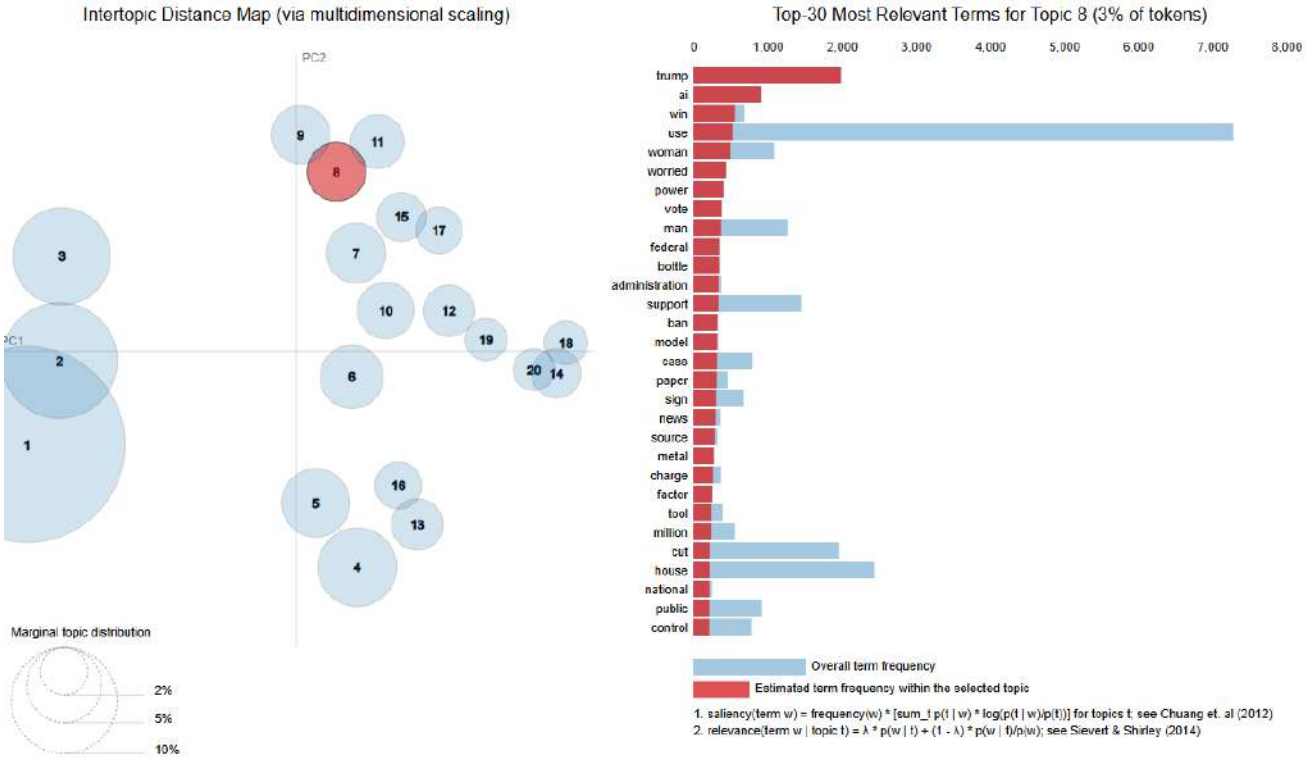

Рисунок 4.1. Топ-слова для однієї з тем, знайденої за допомогою LDA

На основі тем, згенерованих LDA, було здійснено класифікацію кожного повідомлення за домінантною темою — тобто, темою з найвищою ймовірністю для конкретного документа. Це дозволило автоматично призначити кожному тексту тематичну мітку.

Тематичний розподіл усього корпусу (рис. 4.2) виявився неоднорідним, що є очікуваним для текстів із соціальних мереж, де окремі теми домінують у публічному дискурсі. Найбільшою за обсягом виявилася тема 9 (5052 повідомлення), що охоплює загальні емоційні висловлювання та повсякденні рефлексії. Значну кількість повідомлень також отримали теми 1 (1959), 3 (1708), 12 (1366) і 16 (1094), що охоплюють такі напрямки як робота, політика, освіта та повсякденне життя. Натомість теми на кшталт 4 (78 повідомлень), 18 (268), 11 (320) та 8 (302) зустрічаються рідше, що свідчить про їхню специфічність.

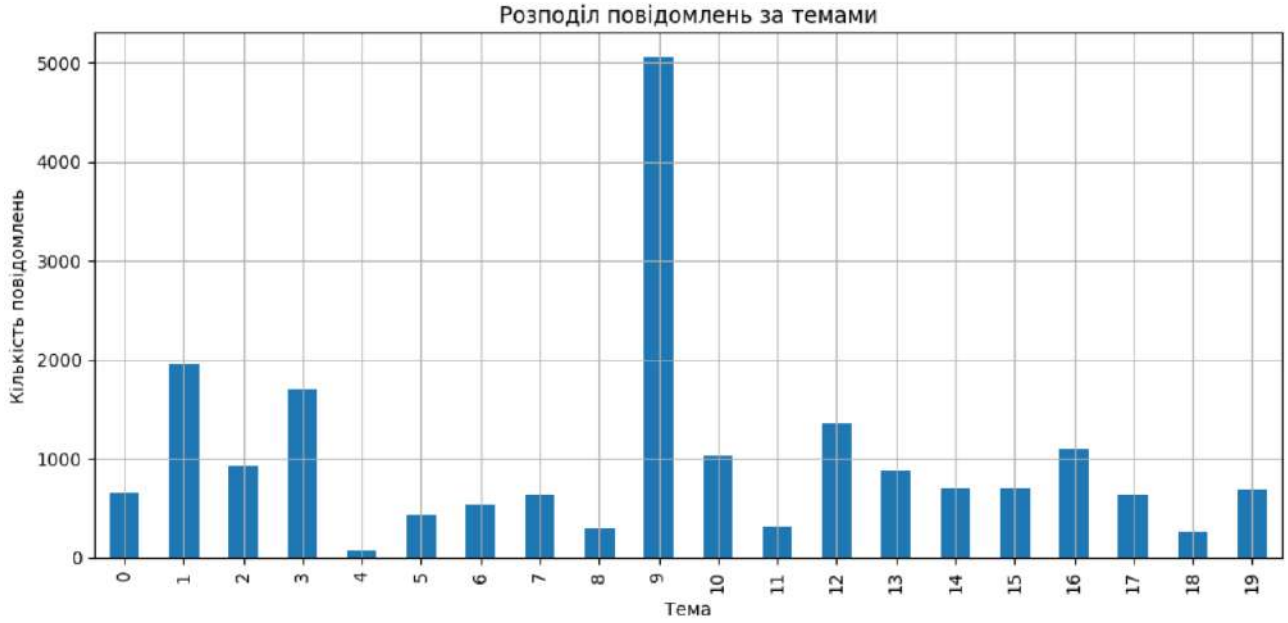

Рисунок 4.2. Кількість повідомлень за темами, знайденими за допомогою LDA

Загалом, модель LDA продемонструвала високу ефективність у виявленні ключових смислових кластерів у текстах соціальних мереж. Незважаючи на нерівномірність розподілу, теми виявилися добре інтерпретованими та змістовно відмежованими, що дозволяє використовувати результати як основу для подальшої автоматичної класифікації повідомлень або побудови тематичних профілів користувачів.

4.2. Семантичне групування повідомлень за допомогою BERT + кластеризації

4.2.1. Кластеризація за допомогою BERT + K-Means

Для виявлення прихованих тем і семантичних зв’язків у великій кількості коротких текстових повідомлень було реалізовано підхід, який поєднує можливості сучасних трансформерних моделей[19] із класичними алгоритмами кластеризації. Зокрема, було використано попередньо натреновану модель Sentence-BERT (all-MiniLM-L6-v2), яка дозволяє отримати компактні і семантично насичені векторні подання (ембедінги) для кожного повідомлення.

На основі чистих текстів було згенеровано ембедінги. Отримані вектори були згруповані за допомогою методу K-Means[20], який є ефективним для кластеризації у векторному просторі. Було обрано 22 кластери, що дозволило досягнути компромісу між деталізацією тем і їх інтерпретованістю. Кількість кластерів підбиралась емпірично на основі змістовності та стабільності топ-слів у кожній групі.

Для кожного кластеру були автоматично визначені ключові слова (рис. 4.3), які найчастіше зустрічались у повідомленнях відповідної групи. Це дало змогу сформулювати узагальнені теми кожного кластеру. Зокрема:

- Кластер 0 (1094 повідомлень) — домашня їжа та рецепти: homemade, chicken, cheese, sauce;
- Кластер 1 (1079) — цифрове та традиційне мистецтво: digital, oil, acrylic, canvas;
- Кластер 2 (551) — родина та виховання дітей: kid, daughter, son, feel;
- Кластер 3 (271) — фітнес і мотивація: fitness, gym, Sunday, victory;
- Кластер 4 (1489) — здоров’я і дослідження: study, cancer, risk, health;
- Кластер 5 (789) — відеоігри та геймерський досвід: game, play, oblivion;
- Кластер 6 (670) — політика й екологія: trump, climate, pollution;
- Кластер 7 (475) — робота та бізнес-середовище: business, job, company, help;
- Кластер 8 (1534) — спорт, змагання: game, world, fan;
- Кластер 9 (1180) — політичні лідери: Trump, Biden, president, win;

- Кластер 10 (700) — питання, відповіді, обговорення: question, post, answer, daily;
- Кластер 11 (882) — освіта: school, student, teacher, education;
- Кластер 12 (946) — кіно та відгуки: movie, film, scene, love;
- Кластер 13 (988) — домашні тварини: dog, cat, pet, vet;
- Кластер 14 (1644) — технології, AI та корпоративна сфера: company, ai, million, pay;
- Кластер 15 (584) — екологія та зміни клімату: climate, change, drought;
- Кластер 16 (605) — догляд за дітьми та немовлятами: old, baby, child, kid;
- Кластер 17 (899) — кар'єра і професійний розвиток: job, career, role, feel;
- Кластер 18 (951) — особисті стосунки і переживання: feel, relationship, friend;
- Кластер 19 (900) — музика та виконавці: song, music, album, artist;
- Кластер 20 (867) — мода і стиль: dress, outfit, style, wedding;
- Кластер 21 (902) — подорожі і місця: trip, travel, flight, place.

```
Кластер 0: homemade, ate, chicken, cheese, sauce, garlic, made
Кластер 1: digital, oil, acrylic, canvas, pencil, paper, watercolor
Кластер 2: kid, feel, day, year, daughter, son, go
Кластер 3: victory, fitness, sunday, welcome, week, gym, let
Кластер 4: study, health, cancer, risk, people, may, find
Кластер 5: game, play, oblivion, year, feel, playing, even
Кластер 6: trump, climate, energy, water, epa, pollution, environmental
Кластер 7: business, job, company, work, need, help, people
Кластер 8: first, game, set, world, year, back, fan
Кластер 9: trump, biden, president, win, house, joe, say
Кластер 10: question, post, sure, search, daily, answer, fitness
Кластер 11: school, student, teacher, education, kid, year, class
Кластер 12: movie, film, best, people, love, scene, good
Кластер 13: dog, cat, pet, vet, day, year, help
Кластер 14: company, ai, million, pay, say, year, day
Кластер 15: climate, change, water, year, study, drought, temperature
Кластер 16: old, kid, year, month, baby, child, day
Кластер 17: job, year, work, career, feel, role, company
Кластер 18: feel, thing, relationship, year, friend, said, even
Кластер 19: song, music, rock, album, love, artist, sound
Кластер 20: dress, outfit, style, look, wear, fit, wedding
Кластер 21: day, trip, go, night, place, travel, flight
```

Рисунок 4.3. Ключові слова за кластерами, знайденими за допомогою BERT + K-means

Кількісний розподіл тем за кластерами подано на рисунку 4.4.

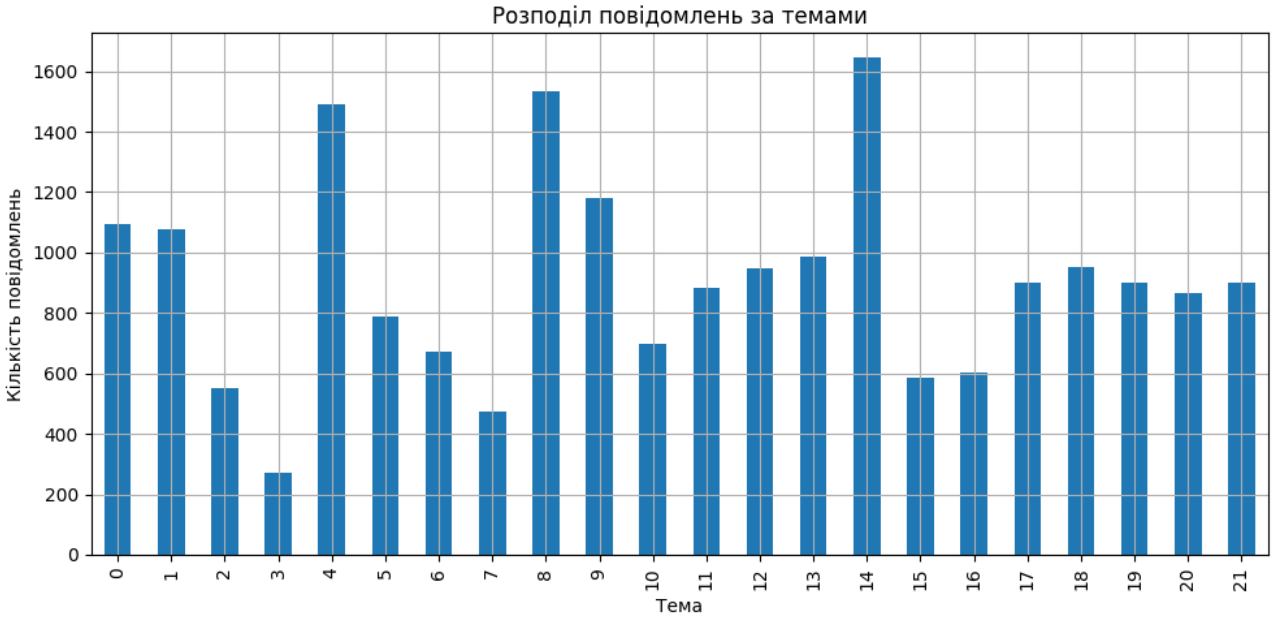

Рисунок 4.4. Кількість повідомлень за темами, знайденими за допомогою BERT + K-means

Кожен кластер охоплює окремий тематичний сегмент, що добре відповідає змісту повідомлень і демонструє здатність BERT-ембедінгів ефективно вловлювати смислову близькість навіть у коротких текстах. Незважаючи на певні перетини між темами, більшість кластерів мають чітке семантичне ядро.

4.2.2. Семантичне групування повідомлень за допомогою BERT + HDBSCAN

У межах цього підходу було реалізовано кластеризацію повідомлень на основі глибоких семантичних подань тексту, отриманих за допомогою моделі BERT, та щільнісного алгоритму кластеризації HDBSCAN[21], який дозволяє автоматично визначати кількість кластерів і виявляти семантично виокремлені групи повідомлень без попереднього маркування.

На першому етапі кожне текстове повідомлення було перетворено у векторне подання за допомогою моделі all-mpnet-base-v2 із бібліотеки SentenceTransformers, яка демонструє високу якість у задачах семантичного порівняння текстів. Далі з метою зменшення розмірності простору та

підвищення ефективності кластеризації було застосовано алгоритм UMAP з параметрами n_neighbors=15, n_components=3, metric='cosine'.

Отримані тривимірні вектори було передано до алгоритму HDBSCAN із параметром min_cluster_size=90, що дозволило виявити 19 тематично узгоджених кластерів, а також частину аномальних повідомлень, яким не було знайдено відповідного кластеру (маркуються як -1). Для підвищення стабільності та повноти результатів було виконано перепризначення outlier-повідомлень до найближчих кластерів за евклідовою відстанню в UMAP-просторі.

Для інтерпретації тем було проведено попередню обробку текстів: видалення стоп-слів, лематизація та частотний аналіз ключових слів у межах кожного кластера. На основі отриманих ключових термінів було визначено тематику кожної групи повідомлень.

Нижче наведено приклади отриманих кластерів і їхніх тематичних описів (таблиця 4.1).

№	**Ключові слова кластера**	**Інтерпретація (тема)**	**Кількість повідомлень**
0	homemade, chicken, cheese, garlic, potato	Домашнє харчування, кулінарія	1021
1	question, post, search, daily, fitness	Щоденні запитання, фітнес-дискусії	742
2	dress, outfit, style, look, wedding	Мода та одяг	1070
3	day, trip, flight, travel, night	Подорожі, поїздки, пригоди	1071
4	digital, oil, acrylic, canvas, pencil	Мистецтво: цифрове і класичне	1091
5	song, music, album, artist, band	Музика, улюблені виконавці	927
6	game, world, fan, player, team	Спортивні події	1120
7	play, oblivion, feel, playing	Геймінг	890
8	movie, film, people, see	Кіно, перегляди, рекомендації	1038
9	dog, cat, pet, vet	Домашні тварини та ветеринарія	1074
10	kid, daughter, son, old	Батьківство, діти, родина	930
11	relationship, boyfriend, friend	Романтичні стосунки	639
12	friend, relationship, together	Дружба, соціальна підтримка	402
13	school, student, teacher, class	Освіта, викладання, учні	1027
14	business, idea, start, company	Стартапи, підприємництво	257
15	job, work, career, role	Професійна діяльність	1134
16	trump, biden, ai, tariff	Політика, штучний інтелект	2390

17	cannabis, marijuana, legal, use	Канабіс, легалізація, дослідження	201
18	study, climate, cancer, people	Наукові дослідження, екологія, здоров'я	2976

Таблиця 4.1. Кластери, отримані за допомогою BERT + HDBSCAN

Таким чином, об'єднання моделей BERT, UMAP та HDBSCAN дало змогу виявити змістовно однорідні групи повідомлень у корпусі. Незважаючи на відсутність використання міток під час кластеризації, алгоритм зміг виділити кластери, що значною мірою відображають наявну тематичну структуру даних. Це свідчить про здатність методу автоматично виявляти смислові патерни у великих обсягах неструктурованого тексту. Отримані кластери можуть слугувати основою для побудови інтерпретованих профілів користувачів або виявлення нових прихованих тем у соціальних медіа.

4.3. Zero-shot класифікація

Zero-shot класифікація[22] — це метод, який дозволяє відносити текстові повідомлення до певних категорій без попереднього навчання моделі на цих конкретних класах. Замість цього модель використовує знання, отримані під час попереднього навчання на завданні розпізнавання природної мови, і здійснює класифікацію, спираючись на логічну схожість між повідомленням і описом категорії.

У цьому дослідженні було обрано 20 тем (див. Розділ 2.1.1), ці категорії були подані моделі як потенційні мітки, а кожне повідомлення оцінювалося на відповідність до них.

Модель joeddav/xlm-roberta-large-xnli була використана через свій високий рівень розуміння контексту. Zero-shot класифікація проводилась батчами по 64 повідомлення, що дало змогу класифікувати повний набір даних на GPU. В результаті найпопулярнішими темами (рис. 4.5) виявилися Travel (2735 повідомлень), Parenting (1872), Business (1473) та Food (1470. Інші широко представлені теми — Technology (1391), Environment (1150) та Education (1108).

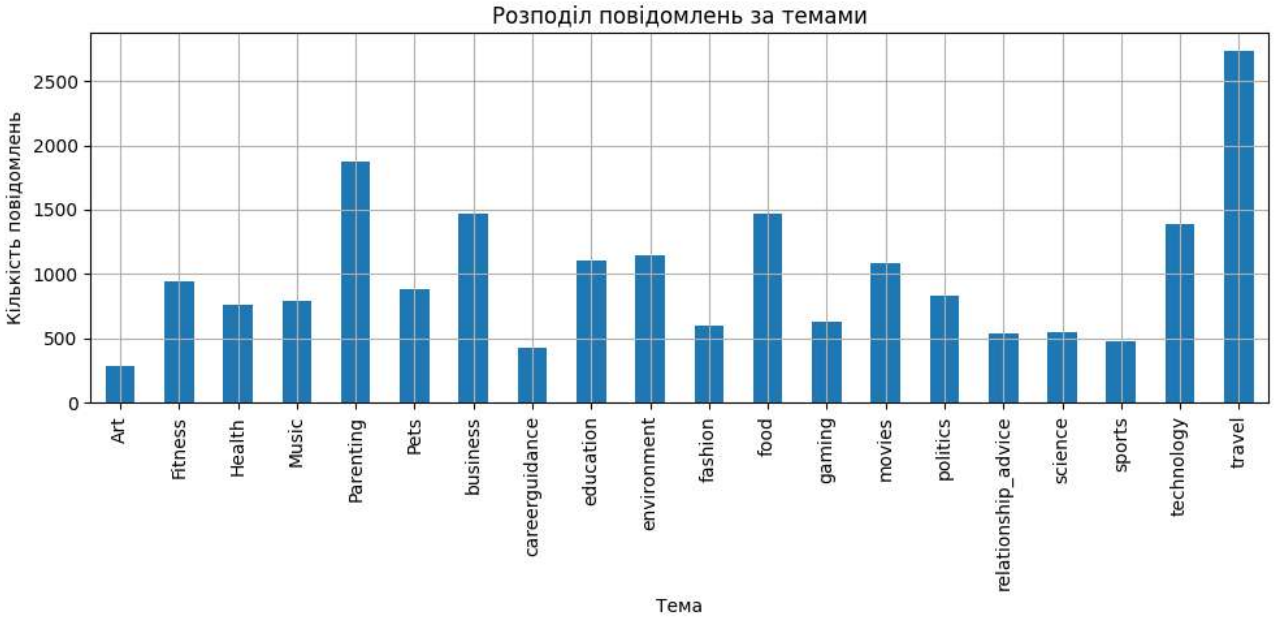

Рисунок 4.5. Кількість повідомлень за темами, zero-shot класифікація

Незважаючи на те, що zero-shot підхід не потребує навчання, він досяг влучності на рівні 52.3% при порівнянні з наявною ручною розміткою. Це є гідним результатом, враховуючи повну відсутність адаптації моделі під специфіку тем. У порівнянні з класичними моделями машинного навчання або тематичним моделюванням (LDA), zero-shot підхід значно виграє в гнучкості, однак поступається в точності, яку демонструють нейромережеві методи з навчанням на специфічному наборі даних.

Висновки до розділу 4

У цьому розділі було реалізовано та порівняно кілька сучасних підходів до тематичної класифікації текстів на основі нейронних мереж та трансформерних моделей. Кожен з методів — LDA, кластеризація BERT-ембедінгів та zero-shot класифікація — продемонстрував свої сильні сторони і межі застосування.

LDA-модель показала високу інтерпретованість тем і дозволила ефективно виявити основні семантичні кластери в корпусі, однак виявила певну чутливість до параметрів і складність у моделюванні коротких текстів.

Кластеризація на основі BERT та алгоритмів K-Means і HDBSCAN виявилася ефективною для автоматичного групування повідомлень за змістом без попереднього маркування. Особливо добре ці підходи працювали у поєднанні з редукцією розмірності UMAP. HDBSCAN дозволив автоматично визначати кількість кластерів і виявляти outlier-повідомлення, що є важливою перевагою для аналізу реальних даних.

Zero-shot класифікація на основі моделі xlm-roberta-large-xnli виявилась зручною для швидкої класифікації без потреби у навчанні. Незважаючи на нижчу точність у порівнянні з іншими методами (52.3% відповідності з ручною розміткою), підхід виявився ефективним для масштабного покриття широкого набору тем.

Узагальнюючи, реалізовані підходи демонструють потенціал гібридних методів: поєднання векторних представлень BERT з кластеризацією або zero-shot логікою дозволяє адаптуватися до різноманітних задач класифікації текстів у соціальних мережах — як із наявною розміткою, так і без неї. Отримані результати можуть бути використані для побудови рекомендаційних систем, аналізу громадської думки або створення профілів користувачів у реальних прикладних задачах.

РОЗДІЛ 5. Порівняння методів тематичної класифікації та оптимальні сценарії їх застосування

У цьому розділі проведено порівняльний аналіз підходів до тематичної класифікації текстів, реалізованих у межах дослідження. Розділ охоплює як класичні алгоритми машинного навчання (наївний баєсівський класифікатор, логістична регресія, метод опорних векторів), так і сучасні підходи (тематичне моделювання LDA, кластеризація BERT-ембедінгів за допомогою K-Means та HDBSCAN), а також zero-shot класифікацію на основі трансформерів без навчання на конкретному наборі тем.

Метою аналізу є не лише оцінка точності моделей, але й виявлення сильних та слабких сторін кожного підходу з точки зору контексту, гнучкості, потреб у навчанні та інтерпретованості результатів. Особливу увагу приділено визначенню сценаріїв, у яких той чи інший метод є найбільш доцільним: для швидкої класифікації нових повідомлень, аналізу неструктурованих даних, або побудови тематичних профілів користувачів.

Результати порівняння дозволяють сформулювати рекомендації щодо вибору алгоритмів залежно від умов задачі, наявності міток, обсягу даних та вимог до точності й інтерпретованості.

5.1. Порівняння класичних моделей машинного навчання

У цьому підрозділі розглянуто ефективність трьох класичних підходів до тематичної класифікації текстів: Multinomial Naive Bayes, Logistic Regression та Support Vector Machine (SVM). Усі моделі було навчено на однорідно підготовленому корпусі повідомлень із соціальних мереж, який було векторизовано за допомогою TF-IDF.

Після попередньої векторизації повідомлень у простір ознак із 10 000–18 000 вимірами кожна модель була навчена й протестована. Оцінку якості проведено за загальними метриками accuracy та F1 macro score, які враховують баланс класифікації в умовах великої кількості тематичних класів:

Модель	Accuracy	F1 Macro
Naive Bayes	0.819	~0.81
Logistic Regression	0.850	~0.84
Support Vector Machine	0.8515	~0.85

Таблиця 5.1. Оцінка якості класифікації класичними моделями

Результати свідчать про те, що Logistic Regression і SVM демонструють вищу точність, ніж Multinomial Naive Bayes, хоча остання модель має перевагу у швидкості та простоті реалізації.

Multinomial Naive Bayes виявився ефективним у темах із чіткою термінологією (наприклад, fitness, pets, art), але мав тенденцію плутати класи з перехресною або абстрактною лексикою (science, technology, business). Врахування біграм і триграм покращило точність, однак модель залишалася обмеженою у врахуванні контексту.

Logistic Regression досягла вищої узагальненої точності. Її здатність враховувати вагові коефіцієнти ознак дозволила краще класифікувати класи з близькою лексикою (наприклад, movies, music, parenting). Водночас, як і всі лінійні моделі, вона не враховувала порядку слів або їхніх глибших взаємозв’язків.

Support Vector Machine (SVM) продемонструвала найвищі показники серед класичних методів. Модель особливо добре справлялась із розділенням компактних тематичних класів (relationship advice, food, gaming). Проте в класах із широкими або семантично перетинаючимися темами (technology, business, science) її точність знижувалась.

Хоча всі три моделі продемонстрували прийнятний рівень якості, їх спільні обмеження вказують на потребу у використанні глибших моделей:

- Вони не враховують контекст або послідовність слів, що є критичним для змістовного аналізу соціальних текстів.
- Вони чутливі до лексичних варіацій: синоніми, переформулювання та сленг можуть знижувати точність.
- Вони не здатні автоматично виділяти семантичні зв’язки між повідомленнями.

У наступному підрозділі буде розглянуто, наскільки нейромережеві та сучасні підходи здатні подолати ці обмеження та підвищити якість тематичної класифікації соціального контенту.

5.2. Порівняння нейромережевих та embedding-підходів

У цьому підрозділі порівнюються три сучасні підходи до тематичного групування текстових повідомлень: Latent Dirichlet Allocation (LDA), BERT + K-Means та BERT + HDBSCAN. Вони реалізують різні принципи обробки тексту, однак мають спільну мету — виявлення латентних тем у великому корпусі неструктурованих повідомлень із соціальних мереж.

BERT + K-Means продемонстрував найвищу відповідність ручному маркуванню тем. Усі 22 сформовані кластери виявились чіткими, інтерпретованими й тематично однорідними. Теми кластерів (наприклад, food, fitness, parenting, gaming, education, travel) збігалися з наданими вручну категоріями, що свідчить про здатність моделі вловлювати глибинні смислові зв’язки. Розподіл повідомлень між кластерами був збалансованим, без надмірного злиття або фрагментації тем.

LDA показала гіршу якість: хоча більшість тем мають інтерпретовані ключові слова, модель продемонструвала сильну нерівномірність у розподілі документів по темах (від 78 до 5000+ повідомлень на тему). Деякі теми виявились надто загальними або змішаними за змістом.

BERT + HDBSCAN виділив 19 тематичних кластерів автоматично, без попереднього завдання їх кількості. Теми кластерів також виявились логічними, а їхня щільність — високою. Однак модель схильна відносити значну кількість повідомлень до аутлайєрів (кластер -1), що потребує подальшого доопрацювання через перепризначення. Якість кластеризації нижча за K-Means, хоча й залишається задовільною.

Складність реалізації та налаштування:

- LDA вимагає складної попередньої обробки текстів (лематизація, очищення, побудова корпусу) і ретельного налаштування гіперпараметрів (num_topics, alpha, eta, кількість проходів).
- BERT + K-Means є найпростішим у реалізації з-поміж embedding-підходів: текст передається у модель Sentence-BERT, далі отримані ембедінги кластеризуються методом K-Means. Єдиний параметр, який потребує підбору — кількість кластерів.
- BERT + HDBSCAN складніший у налаштуванні. Потрібно підібрати параметри UMAP (n_neighbors, n_components) і HDBSCAN (min_cluster_size), а також обробити аутлайєри, які не були автоматично віднесені до кластерів.

Інтерпретованість та щільність кластерів:

- BERT + K-Means забезпечує найкращий баланс між чіткістю, інтерпретованістю та повнотою кластерів. Кожна група має однозначну тему, більшість кластерів легко пояснити за ключовими словами.
- LDA має перевагу у вигляді прямої інтерпретації через список топ-слів для кожної теми, але часто формує теми зі слабкою семантичною відокремленістю.
- BERT + HDBSCAN генерує щільні, вузькі кластери, але інтерпретованість деяких груп може бути зниженою через фрагментацію або наявність схожих тематик у кількох кластерах.

Придатність для побудови профілів користувачів:

- BERT + K-Means — найкращий варіант для побудови узагальнених тематичних профілів. Кожен користувач або документ може бути чітко прив’язаний до однієї теми, що спрощує візуалізацію, аналіз інтересів і побудову рекомендаційних систем.
- LDA — придатна для створення багатотематичних профілів, де один користувач може мати різні теми з відповідними вагами. Такий підхід добре працює в задачах з довшими текстами або при аналізі користувацької активності в динаміці.

- BERT + HDBSCAN доцільний, коли потрібно виділяти вузькі інтереси та аномальні групи, або якщо кількість тем заздалегідь невідома. Підходить для виявлення нішевих тематик.

Отже, найвищу якість тематичного групування в дослідженні продемонстрував підхід BERT + K-Means — завдяки простоті реалізації, високій відповідності ручній розмітці та стабільності результатів. У той час як LDA залишається базовим інструментом із класичної лінійки, а HDBSCAN — потужним інструментом для детального аналізу, саме поєднання BERT-ембедінгів і K-Means є найбільш придатним для задач класифікації коротких повідомлень і побудови зрозумілих тематичних профілів у реальних системах.

5.3. Аналіз zero-shot класифікації

Попри відсутність додаткового навчання, модель досягла 52.3% відповідності ручним міткам — що є гідним результатом для zero-shot підходу. Він виявився особливо ефективним у випадках чітко визначених тем із характерною лексикою (наприклад, Food, Parenting, Gaming).

Порівняно з класичними алгоритмами (TF-IDF + SVM, Logistic Regression, Naive Bayes), zero-shot класифікація значно виграє в гнучкості та універсальності: її можна застосовувати до нових наборів тем без необхідності перенавчання. Утім, за точністю вона поступається як класичним моделям (accuracy до 85%), так і нейромережевим методам із кластеризацією (зокрема BERT + K-Means), які краще адаптовані до структури конкретного корпусу.

Головним недоліком zero-shot підходу є складність розрізнення вузькоспеціалізованих або близьких за значенням тем. Наприклад, модель може змішувати категорії Business і Career Guidance або Health і Science, якщо описові маркери перетинаються. Також модель є чутливою до формулювань назв тем — неправильний вибір лейблу може суттєво знизити якість класифікації.

Zero-shot класифікація є потужним і надзвичайно гнучким інструментом, який дозволяє швидко адаптуватись до нових класифікаційних завдань без етапу

навчання. Вона особливо корисна у випадках обмежених ресурсів, невідомої тематичної структури або потреби в швидкому прототипуванні. Однак для задач, де важлива висока точність, перевагу варто надавати навченим моделям з урахуванням контексту, наприклад, BERT у поєднанні з кластеризацією або нейромережевою класифікацією.

5.4. Порівняльна характеристика методів тематичної класифікації

Метод	Точність / F1	Інтерпретованість	Гнучкість застосування	Необхідність навчання	Складність реалізації
TF-IDF + Naive Bayes	~82% / F1 ≈ 0.81	Висока	Обмежена фіксованими темами	Потребує навчання	Низька
TF-IDF + Logistic Regression	~85% / F1 ≈ 0.84	Висока	Обмежена	Потребує навчання	Середня
TF-IDF + SVM	**~85.2% / F1 ≈ 0.85**	Висока	Обмежена	Потребує навчання	Середня
LDA (Topic Modeling)	~0.53 (Coherence)	Середня (словесні теми)	Гнучка кластеризація	Потребує навчання	Середня
BERT + K-Means	**Найвища відповідність ручній розмітці**	Висока (ключові слова)	Висока	Потребує embedding	Середня
BERT + HDBSCAN	Висока (≈ кластеризація LDA)	Середня (топ-терміни)	Автоматичне виявлення тем	Потребує embedding	Висока
Zero-shot (XLM-RoBERTa)	52.3% відповідності ручній розмітці	Висока (тематичні мітки)	**Найвища (без навчання)**	**Не потребує навчання**	Низька

Таблиця 5.2. Порівняльна таблиця методів тематичної класифікації

Коментарі:

- BERT + K-Means показав найкращу відповідність ручній класифікації при високій чіткості тем.
- Zero-shot класифікація має найвищу гнучкість і не потребує попереднього навчання, але поступається в точності.

- Класичні моделі (SVM, Logistic Regression) мають високу точність, однак не вловлюють контекст та синонімію.
- LDA забезпечує добру інтерпретованість тем і корисна для вивчення загальної структури корпусу, хоча менш точна для класифікації.

5.5. Рекомендації щодо вибору моделі

Результати експериментального порівняння методів тематичної класифікації дозволяють сформулювати практичні рекомендації щодо вибору алгоритмів залежно від конкретної задачі, доступності навчальних даних та вимог до гнучкості або точності.

1. Для швидкої базової класифікації

У випадках, коли потрібна оперативна класифікація великої кількості повідомлень з обмеженими ресурсами або обчислювальними можливостями, доцільно використовувати класичні моделі машинного навчання на основі TF-IDF векторизації:

- TF-IDF + Naive Bayes — найпростіший у реалізації, потребує мінімального налаштування і працює достатньо швидко навіть на великих корпусах.
- TF-IDF + Logistic Regression — забезпечує вищу точність при помірному навантаженні.
- Рекомендується для початкової фільтрації або прототипування систем класифікації.

2. Для високої точності на відомих класах

Коли мета — отримати високу точність класифікації, найкращими варіантами є:

- TF-IDF + SVM — найвища точність серед класичних підходів.
- BERT + K-Means — забезпечує кластеризацію, дуже схожу з ручним маркуванням, особливо ефективна для коротких і семантично насичених повідомлень.

- Рекомендується для промислового розгортання моделей, де класи заздалегідь відомі та добре структуровані.

3. Для аналізу неструктурованих потоків повідомлень

У випадках, коли потік текстових даних не має попередньої розмітки, а теми можуть виникати динамічно, оптимальними є кластеризаційні підходи:

- LDA (Latent Dirichlet Allocation) — дозволяє виявити структуру тем у корпусі й інтерпретувати їх через ключові слова.
- BERT + HDBSCAN — ефективно визначає змістовно однорідні групи повідомлень без необхідності задавати кількість кластерів.
- Підходить для дослідницького аналізу, виявлення нових тем, трендів або поведінкових сегментів.

4. Для мультикатегоріальної класифікації без навчального датасету

Якщо відсутній навчальний корпус, але необхідно класифікувати повідомлення за наперед заданими темами, найкращим варіантом є zero-shot класифікація (XLM-RoBERTa), яка дозволяє застосовувати модель до нових категорій без додаткового навчання.

Підходить для:

- багатомовних сценаріїв;
- швидкого розгортання без анотації;
- адаптації під нові предметні області.

Основне обмеження — нижча точність, особливо на абстрактних або близьких темах.

5. Для побудови профілів користувачів

Завдання побудови семантичних профілів користувачів вимагає методів, здатних виявляти повторювані патерни у текстах конкретного автора:

- BERT + HDBSCAN або K-Means — дозволяють кластеризувати тексти користувача за змістом і виокремити домінантні теми у поведінці.
- LDA — додатково може використовуватись для вивчення тематичної динаміки або визначення інтересів.

- Рекомендується при розробці персоналізованих систем рекомендацій чи профілюванні користувачів.

Загальний висновок:

Жодна модель не є універсальною для всіх задач. Вибір оптимального підходу залежить від:

- наявності або відсутності навчальних даних;
- бажаного балансу між точністю, гнучкістю та швидкістю;
- потреби в інтерпретації або автоматичному виявленні тем.

Висновки до розділу 5

У цьому розділі було проведено всебічне порівняння методів тематичної класифікації текстів, що охоплюють як класичні підходи машинного навчання, так і сучасні нейромережеві та zero-shot рішення. Основні висновки можна сформулювати наступним чином:

- Класичні алгоритми (Naive Bayes, Logistic Regression, SVM) у поєднанні з TF-IDF-векторизацією демонструють гідні результати при низьких обчислювальних витратах. Особливо ефективні для завдань із відомими класами та стабільною лексикою. Найвищу точність серед них показала модель SVM, однак усі класичні підходи поступаються нейромережевим моделям у здатності працювати з контекстом.
- Нейромережеві та embedding-підходи, зокрема BERT + K-Means, , глибоко враховують семантичні подібності між повідомленнями. BERT + HDBSCAN продемонстрував гнучкість у виявленні кластерів без заздалегідь заданої кількості тем, що особливо корисно в аналізі неструктурованих текстових потоків. LDA, попри свою класичну природу, залишається ефективним інструментом для інтерпретованого тематичного моделювання.
- Zero-shot класифікація на базі моделі XLM-RoBERTa дозволяє відносити повідомлення до наперед заданих категорій без навчання, що робить її незамінною в умовах відсутності розміченого датасету. Водночас, вона

поступається за точністю кластеризаторам, навченим на векторних поданнях.

- Вибір методу має базуватись на конкретних завданнях і ресурсах. Для швидкої класифікації — класичні моделі; для високої точності — BERT з кластеризацією; для гнучкої адаптації — zero-shot; для вивчення нових тем або побудови профілів — LDA та HDBSCAN.

Загалом, найкращі результати в цьому дослідженні були досягнуті при використанні BERT + K-Means, що дозволяє рекомендувати цей підхід як базовий варіант для задач тематичної класифікації повідомлень у соціальних мережах. Тим не менш, у реальних прикладних сценаріях доцільно обирати той метод, який найкраще підходить поставленим задачам.

ВИСНОВКИ

Кваліфікаційну роботу присвячено дослідженню, розробці та порівнянню підходів до тематичної класифікації текстових повідомлень у соціальних мережах, з акцентом на їхню ефективність, гнучкість та придатність до автоматизованої обробки великомасштабних неструктурованих даних.

У першому розділі проаналізовано природу соціальних медіа як джерела даних, виявлено основні виклики обробки текстів у таких середовищах (шум, неформальність, відсутність структури), а також розглянуто існуючі методи тематичної класифікації — як класичні, так і нейромережеві.

У другому розділі проведено аналіз предметної області, обґрунтовано вибір тематичних категорій для класифікації, сформульовано задачу та визначено вимоги до системи. Окрему увагу приділено вибору метрик для оцінки якості класифікації та кластеризації.

У третьому розділі реалізовано класичні алгоритми машинного навчання для задачі багатокласової класифікації повідомлень: Naive Bayes, Logistic Regression і SVM. Вони були протестовані з використанням TF-IDF-векторизації. Найвищу якість продемонструвала модель SVM (accuracy ≈ 85.1%), проте всі підходи мали обмеження в роботі з контекстом і синонімією, що зумовлює необхідність нейромережевих підходів.

У четвертому розділі реалізовано сучасні методи класифікації без потреби в ручному маркуванні:

- LDA продемонстрував здатність виявляти інтерпретовані теми, однак мав меншу точність класифікації та нижчу гнучкість;
- BERT + K-Means показав найвищу якість кластеризації, зі значним тематичним збігом із вручну розміченими даними;
- BERT + HDBSCAN виявив змістовно щільні групи без необхідності задавати кількість тем;

- Zero-shot класифікація дозволила виконати повноцінну тематичну розмітку повідомлень без навчального датасету, досягнувши точності ~52.3% — гідний результат для методу без навчання.

У п’ятому розділі проведено комплексне порівняння всіх методів класифікації за точністю, інтерпретованістю, складністю реалізації та сферою застосування. Сформульовано рекомендації щодо вибору методів залежно від задачі:

- для швидкої базової класифікації — Naive Bayes;
- для високої точності на відомих класах — SVM або Logistic Regression;
- для роботи з неструктурованими потоками — BERT + HDBSCAN;
- для zero-shot тематичної розмітки — XLM-RoBERTa;
- для побудови профілів користувачів — LDA або BERT + K-Means.

Узагальнюючи, можна стверджувати, що оптимальним підходом для задачі тематичної класифікації текстів у соціальних мережах є комбінація BERT-ембедінгів із кластеризацією (особливо K-Means), яка забезпечує найкращий баланс між точністю, інтерпретованістю та масштабованістю. Тим не менш, гнучкість zero-shot моделей і простота класичних методів залишають за ними нішу в задачах з обмеженими ресурсами або жорсткими часовими обмеженнями.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Global social media statistics. URL: https://datareportal.com/social-media-users
2. Social Media Usage & Growth Statistics. URL: https://backlinko.com/social-media-users
3. Global social media statistics research summary. URL: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research
4. Доскіч Л. С. Соціальні медіа як джерело дослідницької інформації. Бібліотекознавство. Документознавство. Інформологія. 2023. № 3. С. 105–110.
5. Badry Ali Mustofa, Wawan Laksito Yuly Saptomo. Use of Natural Language Processing in Social Media Text Analysis. Journal of Artificial Intelligence and Engineering Applications. 2025. Vol. 4. No. 2.
6. Aubaid, A. M., Mishra, A. A Rule-Based Approach to Embedding Techniques for Text Document Classification. Applied Sciences, 10(11), 4009. 2020.
7. Das, S., Bhattacharyya, K., Sarkar, S. Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest and SVM on Hate Speech Detection from Twitter. International Research Journal of Innovations in Engineering and Technology, 7(3), 24–28. 2023.
8. Qiang, J., Chen, P., Wang, T., Wu X. Topic Modeling over Short Texts by Incorporating Word Embeddings. 2016.
9. Lu, H., Ehwerhemuepha, L., Rakovski, C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC Medical Research Methodology, 22, 181. 2022.
10. Ma, T., Yao, J.-G., Lin, C.-Y., Zhao, T. Issues with Entailment-based Zero-shot Text Classification. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 786–796. 2021.

11. Digital 2024: Global Overview Report. URL: https://datareportal.com/reports/digital-2024-global-overview-report
12. Srijith, P. K., Hepple, M., Bontcheva, K., Preotiuc-Pietro, D. Sub-story detection in Twitter with hierarchical Dirichlet processes. Information Processing & Management, 53(4), 989–1003. 2017.
13. Соціальна мережа Reddit. URL: https://www.reddit.com
14. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
15. Wang, Sida, and Christopher D. Manning. "Baselines and bigrams: Simple, good sentiment and topic classification." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: *Long Papers-Volume 2*. 2012.
16. Agresti, Alan. Categorical Data Analysis. Wiley, 2002
17. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine Learning, 1995.
18. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3 (2003): 993-1022.
19. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.
20. Jain, Anil K., Murray N. Murty, and Patrick J. Flynn. "Data Clustering: A Review." ACM Computing Surveys, vol. 31, no. 3, 1999, pp. 264-323.
21. Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates." Pacific-Asia Conference on Knowledge Discovery and Data Mining *(PAKDD)*, 2013, pp. 160-172.

22. Liu, Y., Li, S., Zhao, J., Li, Y., Ma, Z., & Chen, G. (2021). Zero-Shot Learning for Natural Language Processing: A Survey. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI).