Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ДОСЛІДЖЕННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДАНИХ У ЗАДАЧАХ ПРИЙНЯТТЯ РІШЕНЬ
(тема)

Виконав:
студент 2 курсу, групи ІНФм-22-2
Алевська А.І.
(прізвище, ініціали)

Спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник доц. Творошенко І.С.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Алевській Анні Ігорівні
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів кластеризації даних у задачах прийняття рішень

затверджена наказом по університету від 3 листопада 2023 року № 1280Ст

2. Термін подання студентом роботи до екзаменаційної комісії 22 грудня 2023 р.

3. Вихідні дані до роботи науково-методична література, тестовий набір даних для проведення дослідження, теоретичні відомості про методи кластеризації, теоретичні відомості про теорію прийняття рішень, перелік використаних програмних засобів: мова програмування Python, середовище розробки IDE PyCharm.

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз існуючих застосунків для кластеризації даних у задачах прийняття рішень.
2. Дослідження механізму кластеризації даних на основі методу спектральної кластеризації (Spectral Clustering).
3. Дослідження механізму кластеризації даних на основі методу LDA.
4. Програмна реалізація методів кластеризації даних у задачах прийняття рішень.
5. Тестування розроблених застосунків та аналіз результатів проведеного дослідження.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) актуальність проблеми кластеризації даних у задачах прийняття рішень, ілюстрація роботи методів, порівняльний аналіз досліджених методів, тестові дані, візуалізація роботи розробленого застосунку.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	03.11.2023	
2	Аналіз завдання, підбір літератури	03.11.23-5.11.23	
3	Аналіз літератури з досліджуваної проблеми	6.11.23-11.11.23	
4	Аналіз технічних засобів	12.11.23-13.11.23	
5	Розробка методу	13.12.23-16.12.23	
6	Програмна реалізація	14.12.23-19.12.23	
7	Оформлення пояснювальної записки	10.12.23-19.12.23	
8	Перевірка на плагіат	01.12.2023	
9	Рецензування	02.12.2023	
10	Підготовка презентації та доповіді	05.12.2023	
11	Занесення роботи в електронний архів	02.01.2024	
12	Попередній захист кваліфікаційної роботи	02.01.2024	

Дата видачі завдання 3 листопада 2023 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Творошенко І.С.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 62 с., 20 рис., 1 табл. 1 дод., 33 джерела.

МЕТОДИ КЛАСТЕРИЗАЦІЇ ДАНИХ, ЗАДАЧІ ПРИЙНЯТТЯ РІШЕНЬ, СПОСОБИ АНАЛІЗУ ДАНИХ, НОРМАЛІЗАЦІЯ ДАНИХ.

Об’єктом дослідження є кластеризація даних в задачах прийняття рішень.

Метою дослідження є вивчення та порівняння методів кластеризації даних у задачах прийняття рішень.

Розглядаються різні методи кластеризації та способи аналізу даних, визначаючи їх основні характеристики та застосування. Загальна важливість досліджень у галузі кластеризації даних полягає в тому, що ці методи дозволяють виявити структуру даних та згрупувати подібні об’єкти разом. Це може бути корисним для багатьох сфер, включаючи машинне навчання, аналіз соціальних мереж, обробку природних мов, медичний аналіз і багато інших.

Проведене дослідження допомагає вирішити проблему ефективності прийняття рішень завдяки створеному оптимальному алгоритму.

DATA CLUSTERING METHODS, DECISION-MAKING PROBLEMS, DATA ANALYSIS METHODS, DATA NORMALIZATION.

The object of research is data clustering in decision-making tasks.

The research aims to study and compare data clustering methods in decision-making tasks.

Different methods of clustering and methods of data analysis are considered, determining their main characteristics and applications. The general importance of the research in the field of data clustering is that these techniques allow us to discover the structure of data and group similar objects together. This can be useful for many fields, including machine learning, social network analysis, natural language processing, medical analysis, and many others.

The conducted research helps to solve the problem of decision-making efficiency thanks to the created optimal algorithm.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

LDA – Latent Dirichlet Allocation (латентне виділення Діріхле)

DBSCAN – Density-Based Spatial Clustering of Applications with Noise (просторова кластеризація застосунків із шумом на основі щільності)

МРТ – магнітно-резонансна томографія

КТ – комп’ютерна томографія

FOREL – формальний елемент

DMSECA – Decision-Making Support for the Evaluation of Clustering Algorithms (підтримка прийняття рішень для оцінки алгоритмів кластеризації)

UCI – University of California, Irvine Machine Learning Repository

MCDM – Multiple-Criteria Decision analysis (багатокритеріальний аналіз рішень)

FBC-Cluster – Fuzzy Behavior Clustering (кластеризація нечіткої поведінки)

HFMCDM – Hesitant Fuzzy Multi Criteria Decision Making (невпевнене прийняття нечітких багатокритеріальних рішень)

FCM – Fuzzy *C*-Means Clustering (нечітка кластеризація *C*-середніх)

HFC – Histogram-Based Fuzzy Clustering (нечітка кластеризація на основі гістограм)

SSE – Sum of Squared Error (сума квадратів похибок)

ВСТУП

В умовах сучасного світу, де доступ до величезних обсягів даних став звичним ділом, завдання прийняття рішень стають все більш складними.

У процесі прийняття рішень, інформація виступає в якості критичного ресурсу, і важливо не лише зібрати ці дані, але також використовувати їх ефективно. Потужним інструментом для аналізу та використання даних є методи класифікації та групування, включаючи процес кластеризації.

Класифікація даних полягає у розділенні набору даних на категорії або класи на підставі конкретних ознак чи характеристик. Цей підхід дозволяє отримувати уявлення, здійснювати прогнози та приймати рішення на основі аналізу даних. Основними методами класифікації є машинне навчання та статистичні підходи, такі як дерева рішень, метод *k*-середніх, нейронні мережі та інші [1].

Сучасні методи кластеризації даних пройшли значний розвиток та розширення протягом останніх років. Ці методи включають такі підходи, як ієрархічна кластеризація, метод *k*-середніх, DBSCAN, агломеративна кластеризація та багато інших. Кожен з них має свої особливості та переваги в конкретних сценаріях використання.

Кластеризація даних грає важливу роль у багатьох сферах. Наприклад, у маркетингу вона дозволяє сегментувати клієнтську базу для більш ефективного спрямування маркетингових зусиль та збільшення лояльності клієнтів. У фінансовому аналізі кластеризація допомагає виявляти аномалії та ризики, сприяючи прийняттю обґрунтованих рішень.

Актуальність дослідження полягає тому, що методи класифікації та групування даних, зокрема кластеризація, є потужними інструментами для аналізу та прийняття рішень у сучасному світі. Розуміння їхніх особливостей та використання відповідно до конкретних завдань може суттєво покращити якість прийнятих рішень та допомогти вирішувати реальні проблеми в різних галузях.

1 АНАЛІЗ ІСНУЮЧИХ ЗАСТОСУНКІВ ДЛЯ КЛАСТЕРИЗАЦІЇ ДАНИХ У ЗАДАЧАХ ПРИЙНЯТТЯ РІШЕНЬ

1.1 Аналіз сучасних застосунків для кластеризації даних у задачах прийняття рішень

Кластеризація даних є важливою складовою багатьох задач прийняття рішень в сучасному світі. Вона допомагає впорядковувати та групувати дані в окремі кластери, що допомагає аналізувати та розуміти закономірності в них. Ця задача стала предметом багатьох досліджень, оскільки вона застосовується в різних областях, таких як сегментація зображень, маркетинг, медицина, соціальні науки, аналіз текстів та інші.

Алгоритми кластеризації дозволяють розділити об’єкти на групи, які схожі за певними критеріями подібності, відомими як кластери. Основною характеристикою цих кластерів є те, що об’єкти в одному кластері схожі один на одного більше, ніж на об’єкти з інших кластерів. Ця класифікація може бути здійснена для великої кількості ознак одночасно.

Можна сформулювати наступні цілі кластеризації:

– розуміння даних шляхом виявлення кластерної структури, що допомагає спростити обробку даних і прийняття рішень, застосовуючи різні методи аналізу для кожного кластера;

– стиснення даних, що дозволяє зменшити розмір вибірки, залишаючи представників найтиповіших кластерів;

– виявлення нетипових об’єктів, які не підпадають під жоден із кластерів;

– використання як попереднього етапу обробки для інших алгоритмів, наприклад, класифікації, які можуть спиратися на заздалегідь визначені кластери;

– підтримка формулювання і перевірки гіпотез на основі аналізу даних.

Кластеризація може мати різні назви в різних контекстах, таких як самонавчання (у визначенні образів), чисельна таксономія (в біології і екології), типологія (в соціальних науках) та розбиття (в теорії графів).

Ця задача отримала велику увагу дослідників, було розроблено багато методів і алгоритмів, однак існують питання, на які ще не знайдено повних відповідей. Однією з найактуальніших проблем у цій області є оцінка результатів та вибір оптимальної кількості кластерів, яка найкраще відображає структуру даних.

Розглянемо деякі з сучасних застосунків кластеризації даних у задачах прийняття рішень:

– маркетинг і сегментація клієнтів: компанії використовують кластеризацію даних для розділення своєї клієнтської бази на групи зі спільними характеристиками. Це допомагає визначити цільові аудиторії, налаштувати маркетингові кампанії та вдосконалити обслуговування клієнтів. Наприклад, Google News використовує кластеризацію статей та новин для групування подій та статей на основі їхньої схожості та тем;

– медична діагностика: у медицині кластеризація даних використовується для аналізу медичних записів та виявлення схожих хвороб чи синдромів. Це допомагає лікарям швидше та точніше ставити діагнози та розробляти індивідуальні схеми лікування. Наприклад, медичні пристрої та програми використовують кластеризацію для аналізу зображень, таких як рентгени, МРТ і сканування КТ, для діагностики та виявлення аномалій;

– фінансовий аналіз: у фінансовому секторі кластеризація даних допомагає виявляти аномалії в фінансових транзакціях, ідентифікувати ризики та визначати інвестиційні можливості;

– соціальні мережі та рекомендації: компанії, які надають послуги у сфері соціальних мереж, використовують кластеризацію даних для підбору рекомендацій, друзів та контенту на основі інтересів користувачів. Наприклад,

Netflix та Amazon Prime Video використовують кластеризацію для рекомендацій фільмів та серіалів користувачам на основі їхнього історії перегляду та вподобань. Spotify використовує кластеризацію для рекомендацій музики, аналізуючи слухачів зі схожими музичними смаками. Facebook використовує кластеризацію для групування користувачів у різні сегменти для налаштування рекламних кампаній та показування змісту. Різні соціальні мережі, такі як Instagram та Pinterest, використовують кластеризацію зображень та вмісту для пошуку та рекомендацій;

– транспорт і логістика: у сфері транспорту і логістики кластеризація даних допомагає оптимізувати маршрути доставки, управляти флотом транспортних засобів та зменшувати витрати на пальне. Наприклад, компанії, такі як FedEx або UPS, використовують кластеризацію для оптимізації маршрутів доставки та управління запасами;

– відстеження аномалій та кібербезпека: кластеризація даних допомагає виявляти аномальну поведінку в комп'ютерних системах та мережах для виявлення потенційних загроз кібербезпеці;

– біологічні дослідження і генетика: у генетиці та біологічних дослідженнях кластеризація даних допомагає групувати гени, білки та інші біологічні об'єкти зі схожими властивостями та функціями;

– аналіз зображень і обробка сигналів: у сферах, як обробка зображень та аналіз сигналів, кластеризація допомагає розпізнавати об'єкти, визначати патерни та виконувати автоматизований аналіз. Наприклад, автопілот Tesla використовує кластеризацію даних з сенсорів та камер для автоматизованого керування автомобілем, а Google Photos використовує кластеризацію для групування фотографій за об'єктами, місцями та обличчями для зручного пошуку.

Застосування кластеризації даних в різних галузях постійно розширюється завдяки розвитку алгоритмів та збільшенню обчислювальних можливостей. Кластеризація даних допомагає покращувати рішення,

зменшувати ризики та оптимізовувати бізнес-процеси у багатьох сферах діяльності [2, 3].

1.2 Класифікація та аналіз існуючих методів для кластеризації даних у задачах прийняття рішень

Алгоритми кластеризації можуть бути простими, такими як алгоритм *k*-середніх, або складнішими, такими як FOREL або алгоритм Ланса-Вільямса. Прості алгоритми, зазвичай, працюють добре лише в обмежених ситуаціях, тоді як складніші алгоритми можуть справлятися з більшим різноманіттям сценаріїв. Однак створення алгоритму, який би був універсальним і працював ідеально у всіх ситуаціях, є досить складною і, в більшості випадків, нерозв'язною задачею.

Для визначення «схожості» об'єктів, спочатку потрібно створити вектор характеристик для кожного об'єкта. Зазвичай це набір числових значень, таких як ріст і вага людини, але також існують алгоритми для роботи з якісними (категорійними) характеристиками. Потім вектори характеристик можуть бути нормалізовані, щоб всі компоненти давали однаковий внесок при обчисленні «відстані». Нормалізація зазвичай обмежує всі значення до певного діапазону, наприклад, [-1, 1] або [0, 1].

На завершальному етапі для кожної пари об'єктів обчислюється «відстань» або ступінь схожості. Існує кілька метрик для обчислення ступеня схожості, таких як:

– Евклідова відстань, ця метрика використовує геометричну відстань у багатовимірному просторі

$$\rho(x, x') = \sqrt{\sum_i^n (x_i - x_i)^2}, \tag{1.1}$$

де ρ − відстань між об'єктами x;

x_i – значення;

i – властивості об’єкта x;

– квадрат Евклідової відстані, ця метрика схильна враховувати більше значущі відстані

$$\rho(x, x') = \sum_i^n (x_i - x_i)^2; \quad (1.2)$$

– відстань Геммінга (Манхеттенська відстань), ця метрика обчислює різницю по кожній координаті і може бути корисною для виявлення різниці в окремих характеристиках

$$\rho(x, x') = \sum_i^n |x_i - x_i|; \quad (1.3)$$

– відстань Чебишева, ця метрика визначається як максимальний модуль різниці між відповідними координатами об’єктів

$$\rho(x, x') = max(|x_i - x_i|); \quad (1.4)$$

– степенева відстань, ця метрика використовується для зважування різниць в окремих координатах

$$\rho(x, x') = \sqrt[r]{\sum_i^n (x_i - x_i)^p}, \quad (1.5)$$

де r і p – параметри, які визначаються користувачем для налаштування впливу різних факторів.

Вибір конкретної метрики залежить від конкретної задачі і вимог дослідника, оскільки він може суттєво вплинути на результати кластеризації.

Отже, вибір метрики відстані є важливою частиною процесу кластеризації і повинен бути обрано з урахуванням конкретних вимог та особливостей задачі [4].

Класифікують такі методи спрямовані на обробку даних:

– неієрархічні методи: ці методи розділяють набір даних на певну кількість окремих кластерів. Їх два основних підходи: перший визначає межі кластерів на основі щільності точок у просторі даних, а другий прагне мінімізувати розбіжність між об’єктами. Алгоритм *k*-середніх є одним з найпоширеніших неієрархічних методів;

– методи за способом аналізу даних: ця категорія поділяється на чіткі та нечіткі алгоритми. Чіткі алгоритми призначені для призначення кожному об’єкту вибірки номера кластера, тобто об’єкт може належати тільки одному кластеру;

– методи за кількістю застосувань алгоритмів кластеризації: в цю категорію входять алгоритми, які можна використовувати для різної кількості кластерів. Наприклад, алгоритм *k*-середніх можна використовувати для будь-якої кількості кластерів, а не тільки для *k* кластерів;

– методи по можливості розширення обсягу оброблюваних даних: деякі методи розроблені з урахуванням можливості розширення обсягу даних, що обробляються. Це може бути важливим, коли ви працюєте з великими обсягами даних або очікуєте, що ваша вибірка буде зростати з часом;

– методи за часом виконання кластеризації: ця категорія охоплює методи, які спроектовані для виконання кластеризації з урахуванням часових обмежень. Тобто вони намагаються знайти компроміс між точністю кластеризації і швидкістю виконання [5].

Неієрархічні методи кластеризації, які базуються на розділенні, представляють собою ітераційні підходи до розбиття початкового набору даних на кластери. У процесі розділення нові кластери формуються до досягнення правила зупинки. Цей тип неієрархічної кластеризації включає в себе розбиття набору даних на окремі кластери. Існують два основних підходи до цього процесу.

Перший підхід полягає в визначенні меж кластерів як найбільш щільних областей в багатовимірному просторі вихідних даних, тобто визначення кластера там, де спостерігається висока «згущеність точок».

Другий підхід спрямований на мінімізацію розбіжності між об’єктами. Найбільш відомим серед неієрархічних методів є алгоритм *k*-середніх, також відомий як метод швидкого кластерного аналізу. Варто відзначити, що для успішного використання цього методу потрібно мати певну гіпотезу про найбільш ймовірну кількість кластерів, тому методи неієрархічної кластеризації, зокрема алгоритм *k*-середніх, не є найкращим вибором для вирішення завдань, де потрібне автоматичне прийняття рішень щодо кількості кластерів [6, 7].

1.3 Особливості задач прийняття рішень

Прийняття рішень – це процес, який включає в себе вибір одного варіанту з ряду альтернатив, які можуть бути реалізовані для вирішення певної проблеми або досягнення певної мети. Цей процес є невід’ємною частиною особистого життя та бізнесу, і від його якості залежить результат і успішність.

Особливості задач прийняття рішень можуть бути різними в залежності від контексту, типу рішення та обставин. Однак деякі загальні особливості варто враховувати:

– неозброєність рішень: більшість рішень приймаються в умовах невизначеності і обмеженої інформації. Це може створювати складнощі і ризики прийняття рішень;

– обмежені ресурси: вирішення задач прийняття рішень може бути обмеженим ресурсами, такими як час, гроші, персонал, інформація тощо;

– альтернативність: все рішення має альтернативи. Важливо розглядати різні можливості та їх наслідки перед вибором оптимального варіанту;

– наслідки і ризики: кожне рішення має свої наслідки та може нести ризики. Важливо визначити потенційні плюси та мінуси кожного варіанту;

– груповий характер: у більшості випадків рішення приймаються в групі. Важливо забезпечити ефективну комунікацію та спільну обговорення;

– часовий фактор: деякі рішення вимагають швидкого прийняття через терміновість ситуації, тоді як інші можуть бути розглянуті більш докладно з більшим строком;

– психологічні аспекти: прийняття рішень може супроводжуватись різними емоціями, страхами, упередженнями та іншими факторами, які можуть впливати на об’єктивність;

– множинні критерії: рішення можуть оцінюватися за різними критеріями, і важливо враховувати всі ці аспекти при прийнятті рішення;

– цикл прийняття рішень: процес прийняття рішень може бути циклічним, особливо в умовах зміни обставин та вхідної інформації;

– системний підхід: у складних задачах прийняття рішень важливо розглядати їх у контексті системи та враховувати взаємодії між елементами системи;

– рішення в умовах невизначеності: прийняття рішень у сучасному світі часто пов’язане з ризиком та невизначеністю, і вимагає врахування цих факторів;

– інформаційні технології: сучасні інформаційні технології та аналітичні інструменти надають можливість збирати, обробляти та аналізувати великі обсяги даних для підтримки прийняття рішень.

Основною метою прийняття рішень є досягнення оптимальних результатів або мінімізація ризиків у вирішенні проблем та досягненні цілей.

Правильний підхід до прийняття рішень включає в себе аналіз інформації, визначення цілей, оцінку альтернатив та їхніх наслідків, а також врахування особливостей контексту і ресурсів [6, 7].

1.4 Аналіз літературних джерел щодо апробації результатів стосовно кластеризації даних у задачах прийняття рішень

У джерелі [1] проведено аналіз актуальних проблем, які існують у сучасному суспільстві, зокрема, проблему низького рівня довіри громадян до обранців. Пропонується один із можливих шляхів вирішення цього питання, а саме використання технології кластеризації. Проведено детальний огляд і аналіз основних методів кластеризації. Сформульовано низку завдань, які потрібно вирішити при застосуванні кластеризації для аналізу груп об’єктів. Однак основною метою є вирішення ключової проблеми – створення можливості для громадян здійснювати обґрунтований вибір під час виборів народних депутатів. Для досягнення цієї мети застосовано високоякісний метод кластеризації, який допоможе зменшити витрати часу, зусиль та ресурсів у процесі прийняття рішень громадянами.

У джерелі [2] розглянуто розробку комп’ютерну систему для аналізу медичних даних за допомогою методу кластерного аналізу. Ця система призначена для підтримки рішень на етапі формування рекомендацій для дослідників, які вибирають певний метод для своїх досліджень. Запропонована схема була протестована на даних, отриманих в результаті клінічних інструментальних досліджень та психологічного обстеження пацієнтів, які страждають від артеріальної гіпертензії.

У джерелі [3] розглядається, що кластерний аналіз виявляється надзвичайно важливим інструментом для аналізу даних, оскільки він допомагає розуміти структуру складних даних, спрощує обробку даних, зменшує їхній обсяг, виявляє нові та нетипові об’єкти і дозволяє формулювати гіпотези. У статті пропонується новий підхід до класифікації об’єктів з урахуванням змін у часі, розроблено інформаційну технологію для підвищення якості кластеризації та проведено практичне впровадження цієї технології в гідрохімічному моніторингу об’єктів із підвищеним

технологічним навантаженням. Це дозволяє зробити процес аналізу та класифікації більш ефективним та інформативним.

У джерелі [4] відображається значний науковий внесок у спільне поле психологічного аналізу прийняття рішень та математичних методів, які використовуються в цій області. Розроблені методи і алгоритми є результатом довгострокових досліджень, під час яких було захищено кілька кандидатських дисертацій і виконано багато наукових робіт. Монографія адресована вченим та практикуючим фахівцям у галузі комп'ютерних наук і прикладної математики, а також економістам, інженерам і адміністраторам, які займаються прийняттям рішень у своїй професійній діяльності. Вона надає можливість оволодіти сучасною методологією та математичними інструментами для покращення процесу прийняття рішень у різних сферах життя.

У джерелі [5] запропоновано єдиний алгоритм ієрархічної кластеризації для прийняття рішень, який спрямований на групування кандидатів та їх законопроектів у подібні категорії, спрощуючи тим самим аналіз даних. Розроблена система спрямована на надання громадянам України можливості зробити інформований вибір серед великої кількості кандидатів на посади депутатів.

У джерелі [6] представлені методи розв'язання різноманітних інтелектуальних завдань, що виникають при аналізі даних і створенні систем штучного інтелекту. Для вирішення завдань з області Data Mining рекомендується використовувати програмні продукти Statistica, MatLab, а також мову програмування Python. У випадку завдань, пов'язаних з розробкою систем штучного інтелекту, пропонується використовувати сучасне програмне середовище Visual Prolog.

У джерелі [7] була розглянута розробка інформаційної технології для підтримки прийняття рішень у ситуаціях невизначеності під час кластерного аналізу. Ця технологія дозволяє одночасно враховувати результати різних функціоналів якості, що дозволяє отримати більш точну оцінку результатів.

Така система була випробувана на практиці за допомогою медичних даних, отриманих під час обстеження хворих на серцеву недостатність.

У джерелі [8] розглядається завдання оцінки алгоритмів обробки великих обсягів даних. Різні алгоритми можуть призводити до різних або навіть протилежних результатів у процесі оцінки, і ця проблема ще не була повністю вивчена. Метою цієї статті є пропозиція схеми для вирішення цієї проблеми шляхом узгодження різних аспектів оцінки алгоритмів кластеризації. Розглянуто модель, відому як «Підтримка прийняття рішень для оцінки алгоритмів кластеризації» (DMSECA), яка дозволяє оцінити алгоритми кластеризації шляхом об'єднання експертних оцінок з метою узгодження відмінностей у їхній ефективності. Модель була перевірена на 20 наборах даних UCI, використовуючи шість алгоритмів кластеризації, дев'ять зовнішніх вимірювань і чотири методи MCDM. Результати показують, що розроблена модель є ефективним інструментом для вибору найбільш підходящих алгоритмів кластеризації для конкретних наборів даних. Крім того, ця модель може узгодити різні або навіть протилежні оцінки для досягнення консенсусу в прийнятті рішень в складному середовищі.

У джерелі [9] розглядає новий метод кластеризації даних під назвою FBC-Cluster. Він базується на нечітких мультиграфах і враховує як структурну, так і атрибутивну подібність вершин. У цьому методі подібність атрибутів визначається за допомогою нечітких T-еквівалентностей серед об'єктів, а структурна подібність враховується за допомогою нового показника подібності, що називається індексом поведінкової подібності. Він використовує замкнуте сусідство в кластерах атрибутів. Результати цього методу включають дві основні категорії кластерів: визначені і можливі, в залежності від порогового рівня β, встановленого для індексу поведінкової подібності. У статті наведено чисельний приклад для демонстрації ефективності цього методу кластеризації. Якість отриманих кластерів також оцінюється за допомогою функцій щільності і ентропії.

У джерелі [10] невизначеність вибору відповідного алгоритму кластеризації моделюється за допомогою невизначеної багатокритеріальної проблеми прийняття рішень (HFMCDM), де деякі алгоритми кластеризації виступають як експерти. Розглянуто нечіткі *C*-середні (FCM) та алгоритми агломеративної кластеризації як дві популярні категорії методів роздільної та ієрархічної кластеризації відповідно. Далі пропонується нова процедура кластеризації, яка базується на невпевнених підходах до прийняття нечітких рішень (HFC), щоб визначити, який з алгоритмів FCM або алгоритмів ієрархічної кластеризації підходить для конкретних даних.

1.5 Постановка задачі дослідження

Таким чином, вирішення задач прийняття рішень за допомогою використання алгоритмів кластеризації даних є актуальним завданням для застосування у широкому спектрі дисциплін. Тому ставиться завдання розробки алгоритму кластеризації, з використанням спільно методу спектральної кластеризації та LDA.

Об’єктом дослідження є кластеризація даних в задачах прийняття рішень.

Метою дослідження є вивчення та порівняння методів кластеризації даних у задачах прийняття рішень.

Для досягнення мети необхідно вирішити такі завдання:

– зібрати та підготувати дані;

– застосувати спектральну кластеризацію;

– побудувати графіки та візуалізації, щоб відобразити кластери та їх структуру;

– визначити ключові ознаки та характеристики кожного кластеру;

– застосувати LDA для кожного кластеру;

– виконати декілька ітерацій, оптимізуючи параметри методів та аналізуючи результати;

– розглянути можливість додавання інших факторів, для покращення кластеризації та аналізу.

2 ОСОБЛИВОСТІ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДАНИХ У ЗАДАЧАХ ПРИЙНЯТТЯ РІШЕНЬ

2.1 Механізм кластеризації даних на основі методу *k*-середніх

Метод *k*-середніх (*K*-Means) – це один із найпопулярніших алгоритмів кластеризації даних. Цей метод допомагає розділити набір даних на *k* кластерів, де кожен кластер містить схожі об’єкти. Механізм роботи методу *k*-середніх включає наступні кроки:

Крок 1. Вибір числа кластерів (*k*). Перший крок – обрати кількість кластерів, на які потрібно розділити дані. Це може бути визначено експертно або за допомогою методів, які оцінюють якість кластеризації, такі як:

– метод ліктя (Elbow Method): він полягає в аналізі графіка залежності суми квадратів внутрішньокластерних відстаней (SSE) від кількості кластерів. «Ліктьова точка» на рисунку 2.1 показує оптимальне значення *k*;

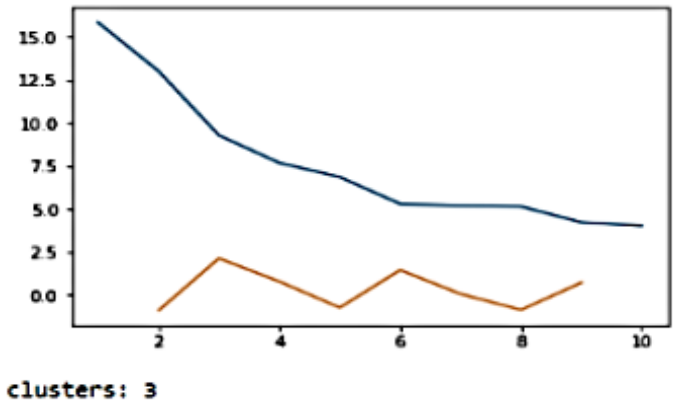

Рисунок 2.1 – Приклад графіку залежності суми квадратів внутрішньокластерних відстаней від кількості кластерів

– метод силуету (Silhouette Method): використовується для обчислення середнього значення силуетів для різних *k*. Вибирається *k*, яке максимізує середній силует (рис. 2.2);

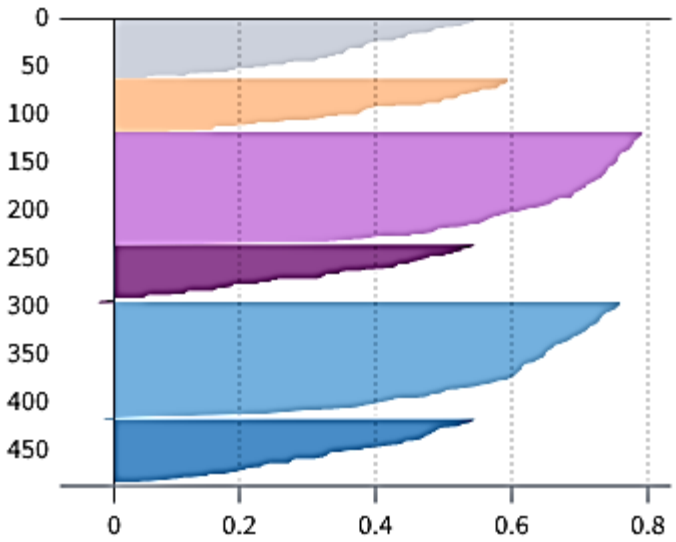

Рисунок 2.2 – Діаграма силуетів

– знання експерта: в деяких випадках, коли у дослідника є експертне розуміння даних, він може заздалегідь визначити, скільки кластерів очікує.

Крок 2. Ініціалізація центроїдів. Центроїди – це точки, які представляють середнє значення об’єктів у кластері. Початкові центроїди для кожного кластера обираються випадковим чином або за допомогою іншого методу ініціалізації. Зазвичай це виконується одним з наступних способів:

– випадковий вибір: початкові центроїди вибираються випадковим чином з діапазону можливих значень;

– вибір об’єктів даних: перші k об’єктів з набору даних використовуються як початкові центроїди;

– *K*-Means++: використовується більш розумний підхід для ініціалізації, який намагається вибрати початкові центроїди, які знаходяться далеко один від одного, щоб поліпшити збіжність алгоритму.

Крок 3. Призначення кожного об’єкта до найближчого кластера. Далі кожен об’єкт у наборі даних призначається до найближчого за відстанню центроїда. Зазвичай використовуються такі метрики відстані, як евклідова відстань, косинусна відстань або Манхеттенська відстань. Це робиться, обчисливши відстань між кожним об’єктом і кожним центроїдом і вибираючи кластер з найменшою відстанню.

Крок 4. Перерахунок центроїдів. Після того, як всі об’єкти призначені до кластерів, центроїди для кожного кластера перераховуються, використовуючи середнє значення всіх об’єктів у цьому кластері. Це означає, що центроїд кожного кластера переміщується в центр мас кластера.

Крок 5. Повторення Кроків 3 та 4. Кроки 3 та 4 повторюються до тих пір, поки центроїди не стабілізуються, тобто не відбуваються зміни або зміни мінімізуються до певного порогу. Це означає, що кластеризація зберігає стабільність, і кластери не змінюються.

Крок 6. Завершення і вивід результату. Коли центроїди стабілізуються, алгоритм завершується, і на виході отримується кластеризація даних. Кожен об’єкт призначений до одного з *k* кластерів.

Крок 7. Оцінка результатів. Наступним кроком може бути оцінка якості кластеризації, наприклад, за допомогою метрик, таких як сума квадратів внутрішньокластерних відстаней (SSE) або зовнішньокластерних відстаней (відстань між кластерами). Ці метрики допомагають визначити ефективність кластеризації, наскільки добре *k*-середніх адаптувався до певних даних.

Ці кроки об’єднуються для виконання ітеративного алгоритму *k*-середніх, який врешті-решт призводить до створення кластерів на основі схожості даних.

Загалом алгоритм можна зобразити у якості блок-схеми, показаної на рисунку 2.3.

Алгоритм *k*-середніх – це ітеративний процес, який пробує знайти найкращі центроїди та розділити дані на *k* кластерів, мінімізуючи внутрішні відстані об’єктів у кластері та максимізуючи відстані між кластерами.

Цей метод добре підходить для великих наборів даних і використовується в різних галузях, включаючи машинне навчання, аналіз даних та графіку [8-10].

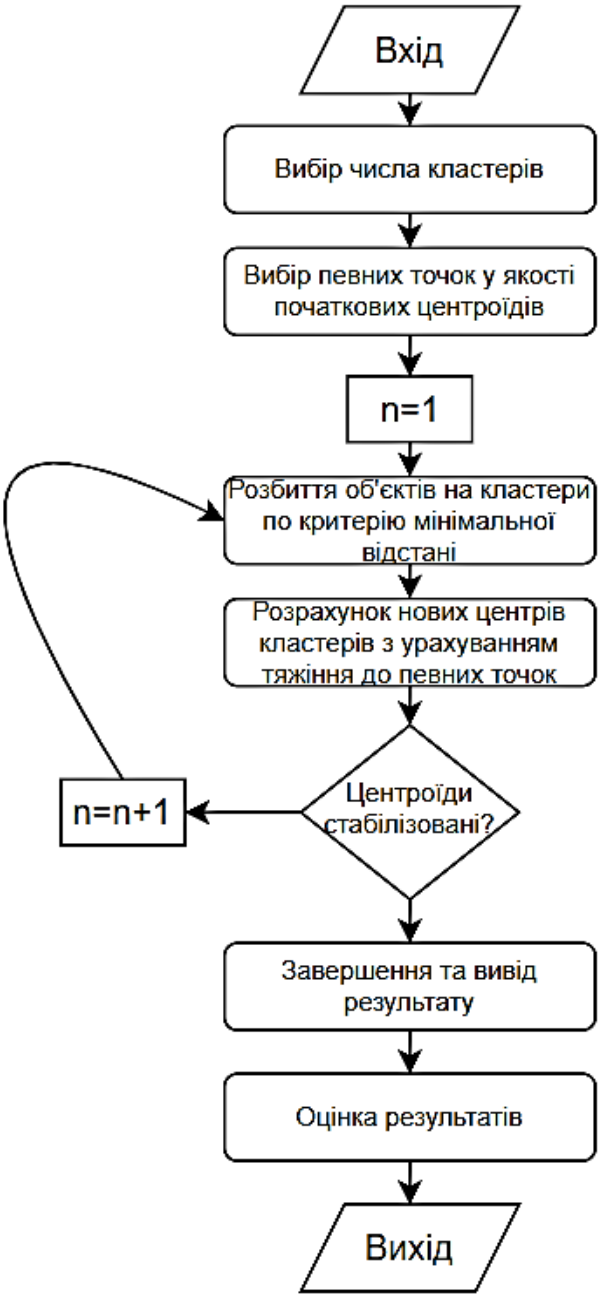

Рисунок 2.3 – Блок-схема алгоритму *k*-середніх

2.2 Механізм кластеризації даних на основі методу агломеративної кластеризації (Hierarchical Clustering)

Агломеративна кластеризація (Hierarchical Clustering) – це метод кластеризації даних, який пробує побудувати ієрархічну структуру кластерів. Цей метод починає з кожного об’єкта, який представляє окремий кластер, і

поступово об’єднує близькі кластери, формуючи більш великі кластери на різних рівнях ієрархії. Основна ідея полягає в тому, щоб мати змогу аналізувати дані на різних рівнях деталізації. Механізм агломеративної кластеризації можна деталізувати наступним чином:

Крок 1. Початкові кластери: на початку кожен об’єкт даних розглядається як окремий кластер. Отже, маємо N початкових кластерів, де N – кількість об’єктів у вихідному наборі даних.

Крок 2. Обчислення матриці відстаней: спочатку обчислюється матриця відстаней між всіма парами об’єктів в наборі даних. Це може бути зроблено за допомогою різних метрик відстані, таких як евклідова відстань, косинусна відстань або кореляційна відстань.

Крок 3. Це означає, що на кожному кроці кількість кластерів зменшується на один, і на виході отримується новий кластер, який об’єднує два попередніх.

Крок 4. Оновлення матриці відстаней: після об’єднання двох кластерів потрібно оновити матрицю відстаней, враховуючи новий кластер. Це може бути зроблено за допомогою різних методів, таких як:

– повторне обчислення відстаней для нового кластера. Можливо, перерахунок відстаней між новим кластером та іншими кластерами;

– метод одного посилання (Single Linkage):

$$d(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} \|x_i - x_j\|, \quad (2.1)$$

де $d(C_i, C_j)$ – відстань між кластерами C_i і C_j;

x_i і x_j – два елементи, між якими розраховується відстань.

Використовується мінімальна відстань між об’єктами двох кластерів;

– метод багатьох посилань (Complete Linkage):

$$d(C_i, C_j) = max_{x_i \in C_i, x_j \in C_j} \|x_i - x_j\|. \quad (2.2)$$

Використовується максимальна відстань між об’єктами двох кластерів;

– метод середнього посилання (Average Linkage):

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} \|x_i - x_j\|. \quad (2.3)$$

Використовується середня відстань між об’єктами двох кластерів;

– метод Уорда (Ward’s Method):

$$d_{ij} = d(\{x_i\}, \{x_j\}) = \|x_i - x_j\|. \quad (2.4)$$

Використовується критерій мінімізації зміни внутрішньокластерних відстаней.

Крок 5. Повторення Кроків 3 та 4: кроки повторюються доти, доки всі об’єкти не об’єднуються в один великий кластер або до досягнення певної структури дерева кластерів.

Крок 6. Побудова дерева кластерів (дендрограми): після завершення алгоритму отримується ієрархічна структура кластерів у вигляді дерева, відомого як дендрограма. Дендрограма відображає ієрархію кластерів і дозволяє аналізувати дані на різних рівнях деталізації.

Крок 7. Вибір кількості кластерів: одним із способів вибору кількості кластерів є зрізання дендрограми на певному рівні. Цей рівень визначає, скільки кластерів дослідник хоче отримати. Дендрограма допомагає визначити оптимальну кількість кластерів відповідно до завдання.

Крок 8. Виведення результату: після вибору кількості кластерів можна використовувати результати агломеративної кластеризації для подальшого аналізу або візуалізації даних.

Узагальнити роботу алгоритму можна наступною блок-схемою на рисунку 2.4.

Рисунок 2.4 – Блок-схема роботи алгоритму агломеративної кластеризації

Агломеративна кластеризація – це потужний метод для виявлення структури в даних, особливо коли важко передбачити кількість кластерів апріорі. Вона широко використовується в багатьох галузях, включаючи біологію, соціальні науки, медицину, та інші [8-10].

2.3 Механізм кластеризації даних на основі методу спектральної кластеризації (Spectral Clustering)

Спектральна кластеризація (Spectral Clustering) – це метод кластеризації даних, який використовує інформацію з власних векторів та власних значень матриці схожостей для розділення даних на кластери. Цей метод особливо ефективний для даних, які мають складну нееліптичну форму, але він вимагає обчислення і аналізу великої матриці схожостей та власних векторів, тому він може бути витратний за обчисленнями.

Детальний механізм спектральної кластеризації (рис. 2.5):

Крок 1. Побудова матриці схожостей: спершу створюється матриця схожостей, яка відображає схожість між об’єктами даних. Зазвичай це є симетрична матриця, де кожен елемент показує ступінь схожості між відповідними об’єктами. Можна використовувати різні метрики схожості, такі як евклідова відстань, косинусна схожість або якась інша відповідно до завдання.

Крок 2. Створення нормалізованої матриці схожостей: за допомогою матриці схожостей створюється нормалізована матриця. Один із способів це зробити – це використовувати «симетричну нормалізовану матрицю схожостей», в якій кожен елемент ділиться на суму схожостей цього об’єкта до всіх інших об’єктів.

Крок 3. Побудова Графу схожостей: нормалізована матриця схожостей може бути розглянута як матриця ваг, і на її основі будується граф схожостей. В цьому графі кожен об’єкт відповідає вузлу, а ребра між вузлами представляють ступінь схожості між об’єктами.

Крок 4. Розрахунок Лапласіана графа: Лапласіан графа – це матриця, яка визначається на основі матриці схожостей та нормалізованої матриці схожостей. Він включає інформацію про зв’язки між вузлами графа.

Крок 5. Обчислення власних векторів та власних значень Лапласіана: далі, розв’язується задача власних значень і власних векторів для матриці

Лапласіана графа. Власні вектори представляють характеристики графа, і вони використовуються для подальшого розділення даних на кластери.

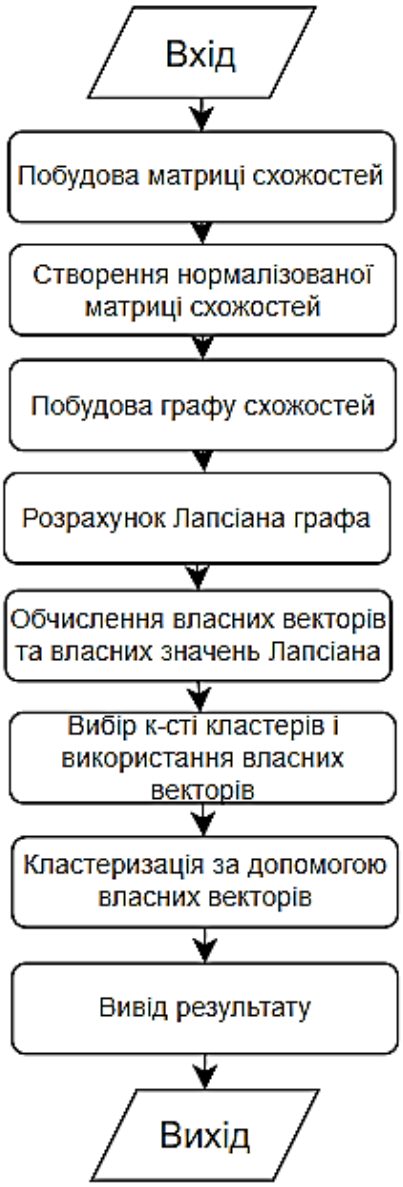

Рисунок 2.5 – Блок-схема алгоритму спектральної кластеризації

Крок 6. Вибір кількості кластерів і використання власних векторів: вибір кількості кластерів зазвичай визначається експертно або за допомогою методів, які аналізують власні значення, такі як метод ліктя. Потім використовуються перші k власних векторів, де k – кількість кластерів.

Крок 7. Кластеризація за допомогою власних векторів: власні вектори використовуються як функції для переведення даних в новий простір, де проводиться кластеризація, зазвичай за допомогою алгоритмів кластеризації, таких як k-середніх або ієрархічна кластеризація.

Крок 8. Виведення результату: результатом спектральної кластеризації є кластери об’єктів даних, які були розділені на основі їхньої структури схожості. Кожен об’єкт належить до одного з кластерів.

Спектральна кластеризація може бути дуже ефективною для даних зі складною структурою та нерегулярною формою кластерів. Вона дозволяє виявити глобальні та локальні структури у даних і використовується в багатьох галузях, включаючи аналіз зображень, сегментацію тексту та біологічну класифікацію [11-14].

2.4 Механізм кластеризації даних на основі методу LDA

Метод Latent Dirichlet Allocation (LDA) – це статистичний метод, який використовується для кластеризації текстових даних на теми. LDA є типом тематичної моделі, яка намагається виявити латентні теми, які відповідають регулярним візуальним або семантичним паттернам в текстах. Основна ідея полягає в тому, що кожен документ представляється як сума декількох тем, і кластеризація відбувається на основі цих тем.

Механізм LDA у більш детальному розгляді (рис. 2.6):

Крок 1. Підготовка текстових даних: перший крок – це підготовка текстових даних. Це може включати в себе очищення тексту від зайвих символів, токенізацію (розділення тексту на окремі слова або терміни), видалення стоп-слів (загальних слів, які не несуть значущої інформації, таких як «і», «у», «до»), і лематизацію (перетворення слів до їхньої базової форми).

Крок 2. Побудова словників: для кожного тексту будується словник, де кожне слово або термін має свій унікальний ідентифікатор. З цими словниками

створюються вектори документів, які показують, які слова присутні в кожному документі та скільки разів кожне слово використовується в документі.

Рисунок 2.6 – Блок-схема алгоритму LDA

Крок 3. Вибір гіперпараметрів LDA: для моделі LDA необхідно вибрати гіперпараметри, такі як кількість тем та гіперпараметри розподілу Діріхле. Гіперпараметри визначають, скільки кожний документ має теми, і скільки кожна тема має слів.

Крок 4. Оцінка моделі LDA: перед використанням LDA, модель зазвичай оцінюється на даних для підбору найкращих гіперпараметрів. Це включає в

себе використання перехресної перевірки та метрик, таких як перплексія, для оцінки якості моделі.

Крок 5. Навчання LDA: навчання LDA включає в себе використання методу вищого порядку, яким є розподіл Діріхле, для оцінки того, які слова пов’язані з якими темами. Метою є знайти найкращі значення розподілу Діріхле, які найкраще пояснюють дані.

Крок 6. Кластеризація даних на основі тем LDA: після навчання LDA кожен документ представляється як сума тем, і кластеризація документів відбувається на основі того, як сильно кожен документ пов’язаний з різними темами. Можна використовувати методи кластеризації, такі як *k*-середніх або ієрархічна кластеризація, для розділення документів на кластери на основі їхнього векторного представлення тем.

Крок 7. Аналіз результатів і інтерпретація тем: після кластеризації можна проаналізувати результати та інтерпретувати знайдені теми. Це може включати в себе аналіз найбільш репрезентативних слів для кожної теми, а також аналіз документів, які входять до кожного кластера.

Метод LDA допомагає розкрити латентні теми в текстових даних і кластеризувати їх на основі цих тем. Він широко використовується в аналізі текстів, тематичному моделюванні та видачі рекомендацій на основі змісту документів [15-17].

2.5 Методика кластеризації даних у задачах прийняття рішень

Методика кластеризації даних грає важливу роль у задачах прийняття рішень, оскільки дозволяє групувати схожі дані та виділяти закономірності у великих об’ємах інформації.

Різні методі кластеризації можуть бути використані для певних задач, вони мають різні характеристики та ефективність (табл. 2.1). Залежно від завдання та властивостей даних, вибір конкретного методу кластеризації може

бути обумовлений різними факторами. Важливо експериментувати та проводити оцінку результатів для вибору оптимального методу для конкретного випадку використання.

Таблиця 2.1 – Порівняльний аналіз досліджених методів кластеризації

Характеристика	***k*-середніх (K-Means)**	**Агломеративна кластеризація**	**Спектральна кластеризація**	**LDA**
Ефективність кластеризації	Добре працює для кругових та сферичних кластерів, але не ефективний для некругових та нелінійних.	Добре працює для різних форм кластерів, але може бути обчислювально витратною для великих обсягів даних.	Ефективна для розрізнення складних форм кластерів та працює добре при нелінійних залежностях.	Призначений для тематичної моделі, не для точкової кластеризації.
Спроможність роботи з великими обсягами даних	Добре підходить для великих обсягів даних, але чутливий до початкового вибору центроїдів.	Може бути витратною для великих обсягів даних, особливо при використанні методу «повного з’єднання».	Зазвичай витратна за часом та пам’яттю для великих обсягів даних.	Зазвичай ефективний для великих обсягів текстових даних.

Продовження таблиці 2.1

Характеристика	***k*-середніх (*K*-Means)**	**Агломеративна кластеризація**	**Спектральна кластеризація**	**LDA**
Взаємодія з формою та розміром кластерів	Чутливий до розміру та форми кластерів, ефективний лише для круглих форм.	Добре працює з різними формами та розмірами кластерів.	Добре розрізняє складні форми кластерів, але може бути чутливою до параметрів.	Не призначений для точкової кластеризації.
Необхідність задання кількості кластерів	Потребує попередньо визначеної кількості кластерів.	Має гнучкість у визначенні кількості кластерів, але може вимагати вибору оптимального рівня відсічення.	Має гнучкість у визначенні кількості кластерів, але параметри важливі.	Має гнучкість у визначенні кількості тем.
Обробка категоріальних даних	Працює з числовими даними, не ефективний для категоріальних.	Може працювати з різними типами даних, включаючи категоріальні.	Працює з числовими даними, може потребувати додаткових перетворень для категоріальних.	Призначений для роботи з текстовою інформацією.

Продовження таблиці 2.1

Характеристика	**k-середніх (K-Means)**	**Агломеративна кластеризація**	**Спектральна кластеризація**	**LDA**
Часова складність алгоритму	$O(k * n * l * d)$, де k – кількість кластерів; n – кількість об’єктів; l – кількість ітерацій; d – кількість ознак.	$O(n^2 * \log n)$, або $O(n^3)$, залежно від використовуваного методу (наприклад, «повне з’єднання» або «одиничне з’єднання»).	Зазвичай $O(n^2 * \log n)$, або $O(n^3)$, залежно від методу та використання матриць суміжності та Лапласіана.	Зазвичай $O(I * T * * D * W)$, де I – кількість ітерацій; T – кількість тем, D – кількість документів; W – середня кількість слів у документі.

Нижче розглянуто, як методика кластеризації даних може бути використана для прийняття рішень:

– пошук структури даних: кластеризація даних дозволяє виявити структуру у великому наборі даних. Це може бути корисно при аналізі даних і розумінні, як об’єкти пов’язані один з одним. Наприклад, в бізнес-аналізі, це може допомогти зрозуміти, які групи клієнтів мають схожі властивості і як це впливає на їхню поведінку;

– кластеризація для сегментації аудиторії: у маркетингу кластеризація допомагає сегментувати аудиторію на підгрупи зі схожими інтересами та потребами. Наприклад, кластеризація клієнтів у роздрібній торгівлі допомагає розуміти, які товари чи послуги цікавлять різні групи споживачів;

– виявлення аномалій: кластеризація може бути використана для виявлення аномалій або несподіваних груп об’єктів. Наприклад, в детекції шахрайства кредитних карт, аномальні транзакції можуть бути виявлені, аналізуючи кластери звичайних та аномальних транзакцій;

– покращення прийняття рішень в медицині: у медицині кластеризація може бути використана для групування пацієнтів за схожими медичними історіями та характеристиками. Це може допомогти лікарям в уточненні діагнозів та розробці індивідуальних планів лікування;

– вибір стратегій управління: у сфері управління, наприклад, у роботі містами, кластеризація допомагає управлінцям визначити, які райони або групи населення мають схожі потреби та вимоги. Це важливо при прийнятті рішень щодо виділення ресурсів, плануванні інфраструктури, а також в справах оборони і безпеки;

– оптимізація рекомендаційних систем: в інтернет-сервісах кластеризація даних може бути використана для покращення рекомендаційних систем. Користувачі можуть бути поділені на групи зі схожими інтересами, і рекомендації можуть бути зроблені на основі попередніх дій інших користувачів у тому ж кластері;

– оцінка ризиків імовірних подій: кластеризація може допомогти в оцінці ризиків та прийнятті рішень у фінансовій галузі, страхуванні, та інших сферах, де важлива оцінка ймовірності різних подій.

Усі ці застосування використовують методику кластеризації даних для отримання корисної інформації і підтримки прийняття рішень. Кластеризація дозволяє виявляти залежності та групи в даних, що може бути важливим для вирішення різних завдань у різних галузях [18, 19].

3 ДОСЛІДЖЕННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДАНИХ У ЗАДАЧАХ ПРИЙНЯТТЯ РІШЕНЬ

3.1 Вибір інструментальних засобів для реалізації методів кластеризації даних у задачах прийняття рішень

У рамках кваліфікаційної роботи було об’єднано та подано методи LDA та спектральної кластеризації даних у задачах прийняття рішень. Для реалізації було обрано мову програмування Python та IDE PyCharm. Це обумовлено тим, що Python часто вважається більш зручним та ширше використовується в галузі науки про дані та машинного навчання.

Кластеризація – це процес поділу різних частин даних на основі загальних характеристик. Python пропонує багато корисних інструментів для виконання кластерного аналізу. Найкращий інструмент для використання залежить від поточної проблеми та типу доступних даних. Існує три широко використовувані методи формування кластерів у Python: кластеризація *k*-середніх, моделі сумішей Гауса та спектральна кластеризація. Для відносно малорозмірних завдань (щонайбільше кілька десятків вхідних даних), таких як ідентифікація окремих груп споживачів, кластеризація *k*-середніх є чудовим вибором. Для більш складних завдань, таких як, наприклад, виявлення нелегальної діяльності на ринку, краще підійде більш надійна та гнучка модель, така як модель сумішей Гуасса. Нарешті, для задач великої розмірності з потенційно тисячами вхідних даних найкращим варіантом є спектральна кластеризація [20-23].

Для дослідження обраних методів кластеризації ефективність використання мови програмування Python також обумовлено наступними перевагами:

– широкий спектр бібліотек для обробки даних та машинного навчання:

1) scikit-learn: ця бібліотека надає багато інструментів для кластеризації, включаючи Spectral Clustering та KMeans. Також, вона має інші корисні функції для обробки даних та використання моделей машинного навчання;

2) gensim: для роботи з LDA та тематичною моделлю;

– популярність та активна спільнота: Python має широку та активну спільноту розробників, яка підтримує багато бібліотек та фреймворків. Це означає, що є змога швидко знайти допомогу, ресурси та оновлення;

– прототипування та читабельність коду: Python добре підходить для прототипування завдяки своєму простому та читабельному синтаксису. Можна швидко реалізовувати та тестувати ідеї;

– ефективність роботи з текстовими даними: Python має багато бібліотек для роботи з текстовими даними, такими як pandas для обробки даних та nltk або spaCy для обробки тексту;

– легка інтеграція з іншими інструментами: Python є популярною мовою для наукового стеку, який включає багато інших інструментів для візуалізації, аналізу даних та інше. Наприклад, можливо використовувати бібліотеку matplotlib для графічного відображення результатів.

3.2 Етапи програмної реалізації методів кластеризації даних у задачах прийняття рішень

Для реалізації поставленої задачі необхідно виконати наступні завдання:

Крок 1. Збір та підготовка даних:

– збір даних з соціальних мереж, включаючи інформацію про користувачів, рекламні дописи, активність та текстову інформацію (рис. 3.1);

– попередня обробка даних, така як видалення дублікатів, очищення текстової інформації від стоп-слів і символів пунктуації.

	# adid	adtext	# clicks	# impressions	age	creationdate
1	374	Join us because we care. Black matter	0	137	18 - 65+	06/10/15 02:59:53 AM PDT
2	655	NOT EVERY BOY WANTS TO BE A SOLDIER.	35	452	18 - 65+	06/23/15 07:04:01 AM PDT
3	664	"People can tolerate two homosexuals	26	374	18 - 65+	06/23/15 07:02:40 AM PDT
4	79	?????? ??? ????? ? ?????????	0	31	18 - 65+	06/09/15 03:50:21 AM PDT
5	325	California... knows how to party Cali	4	326	18 - 65+	06/10/15 07:34:52 AM PDT
6	326	Since 2010, over 350 of our lives hav	517	1478	18 - 65+	06/12/15 03:13:16 AM PDT
7	327	'Just like Trayvon Martin, race matte	7	125	18 - 65+	06/11/15 06:51:30 AM PDT
8	328	Race war started by Texas teacher A T	17	168	18 - 65+	06/11/15 07:03:58 AM PDT
9	329	The image of 1938 shows several Afric	18	482	18 - 65+	06/15/15 07:21:33 AM PDT
10	330	American Racists On The Road T	24	524	18 - 65+	06/15/15 07:22:00 AM PDT
11	331	"Free Figure's Black Power Rally at V	43	764	18 - 65+	06/15/15 07:21:47 AM PDT
12	332	A woman pretended Afro-American to ga	47	676	18 - 65+	06/15/15 07:22:20 AM PDT
13	333	2Pac believed in Fight Tupac Shakur w	47	1075	18 - 65+	06/16/15 07:36:57 AM PDT
14	334	Today we celebrate the legendary rapp	10	153	18 - 65+	06/16/15 07:37:14 AM PDT
15	335	It is an American history. African-Am	26	476	18 - 65+	06/16/15 08:20:31 AM PDT
16	336	Americans so much effort to hate each	42	745	18 - 65+	06/17/15 07:41:34 AM PDT
17	337	No national outrage for an incident a	23	746	18 - 65+	06/17/15 07:41:07 AM PDT
18	338	A 13-year-old child was tasered by co	97	1290	18 - 65+	06/17/15 07:41:46 AM PDT
19	339	Only 9 days this year nobody was kill	55	1005	18 - 65+	06/17/15 07:42:02 AM PDT
20	340	In early 1900s this Black Americana C	21	844	18 - 65+	06/17/15 07:45:33 AM PDT
21	341	Sadness and shocking tragedy at histo	274	5072	18 - 65+	06/18/15 04:36:59 AM PDT
22	342	Rachel Dolezal is wearing the other p	46	866	18 - 65+	06/18/15 02:50:25 AM PDT

Рисунок 3.1 – Приклад даних, зібраних з соціальної мережі Facebook

Виконання попередньої обробки даних для csv-файлу включає в себе кілька етапів, що реалізуються з допомогою бібліотек pandas та nltk. Результат обробки представлений на рисунку 3.2.

Крок 2. Застосування спектральної кластеризації:

– використання спектральної кластеризації для грубої кластеризації рекламних дописів на основі взаємодій з ними користувачів та інших ознак;

– визначення кількості кластерів або використання методів визначення оптимальної кількості кластерів.

Крок 3. Візуалізація результатів:

– побудова графіків та візуалізації, для відображення кластерів та їх структури (рис. 3.3);

– визначення ключових ознак та характеристик кожного кластеру.

	adid	adtext	clicks	impressions	age	creationdate
1	374	join us care black matters	1	137	18 - 65+	06/10/15 02:59:53 AM PDT
2	655	every boy wants soldier beautiful mes	35	452	18 - 65+	06/23/15 07:04:01 AM PDT
3	664	people tolerate two homosexuals see l	26	374	18 - 65+	06/23/15 07:02:40 AM PDT
4	79	No data.	1	31	18 - 65+	06/09/15 03:50:21 AM PDT
5	325	california knows party california kno	4	326	18 - 65+	06/10/15 07:34:52 AM PDT
6	326	since 2010 350 lives taken hands poli	517	1478	18 - 65+	06/12/15 03:13:16 AM PDT
7	327	like trayvon martin race mattered ama	7	125	18 - 65+	06/11/15 06:51:30 AM PDT
8	328	race war started texas teacher texas	17	168	18 - 65+	06/11/15 07:03:58 AM PDT
9	329	image 1938 shows several african amer	18	482	18 - 65+	06/15/15 07:21:33 AM PDT
10	330	american racists road racists group a	24	524	18 - 65+	06/15/15 07:22:00 AM PDT
11	331	free figures black power rally vcu	43	764	18 - 65+	06/15/15 07:21:47 AM PDT
12	332	woman pretended afroamerican gain con	47	676	18 - 65+	06/15/15 07:22:20 AM PDT
13	333	2pac believed fight tupac shakur inde	47	1075	18 - 65+	06/16/15 07:36:57 AM PDT
14	334	today celebrate legendary rapper ever	10	153	18 - 65+	06/16/15 07:37:14 AM PDT
15	335	american history africanamerican citi	26	476	18 - 65+	06/16/15 08:20:31 AM PDT
16	336	americans much effort hate instead pu	42	745	18 - 65+	06/17/15 07:41:34 AM PDT
17	337	national outrage incident pool party	23	746	18 - 65+	06/17/15 07:41:07 AM PDT
18	338	13yearold child tasered cop skateboar	97	1290	18 - 65+	06/17/15 07:41:46 AM PDT
19	339	9 days year nobody killed police acco	55	1005	18 - 65+	06/17/15 07:42:02 AM PDT
20	340	early 1900s black americana chocolate	21	844	18 - 65+	06/17/15 07:45:33 AM PDT
21	341	sadness shocking tragedy historically	274	5072	18 - 65+	06/18/15 04:36:59 AM PDT
22	342	rachel dolezal wearing persons shoes	46	866	18 - 65+	06/18/15 02:50:25 AM PDT

Рисунок 3.2 – Дані після обробки

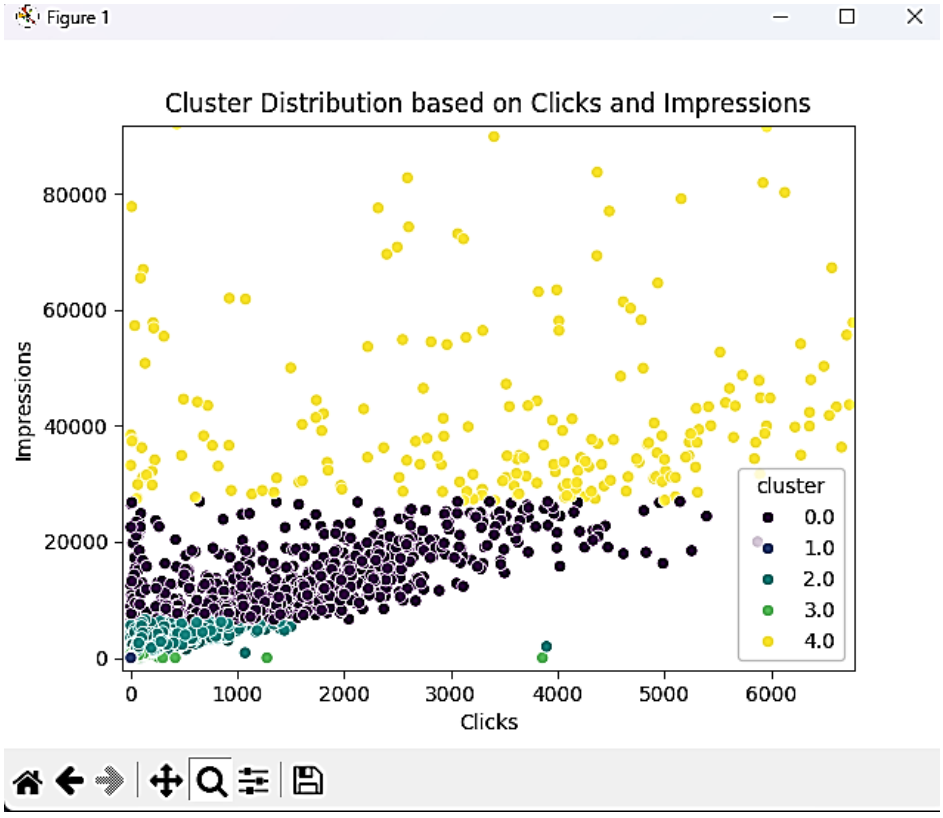

Рисунок 3.3 – Візуалізація створених кластерів

Крок 4. Застосування LDA для кожного кластеру:

– розділення реклами на кластери, що були знайдені за допомогою спектральної кластеризації за запис вихідних даних у файл (рис. 3.4);

– для кожного кластеру виконання аналізу тематичних патернів за допомогою LDA на текстових даних (рис. 3.5).

	AdText	# clicks	# Impressions	# cluster
1	join us care black matters	1,0	137,0	3,0
2	every boy wants soldier beautiful mes	35,0	452,0	3,0
3	people tolerate two homosexuals see l	26,0	374,0	3,0
4	No data.	1,0	31,0	3,0
5	california knows party california kno	4,0	326,0	3,0
6	since 2010 350 lives taken hands poli	517,0	1478,0	2,0
7	like trayvon martin race mattered ama	7,0	125,0	3,0
8	race war started texas teacher texas	17,0	168,0	3,0
9	image 1938 shows several african amer	18,0	482,0	3,0
10	american racists road racists group a	24,0	524,0	3,0
11	free figures black power rally vcu	43,0	764,0	3,0
12	woman pretended afroamerican gain con	47,0	676,0	3,0

Рисунок 3.4 – Розділення реклами на кластери

```
Cluster 0 Topics:
Topic 0: 0.026*"black" + 0.007*"u" + 0.007*"matter" + 0.007*"people" + 0.006*"african" + 0.006*"police" + 0.005*"community" + 0.005*"stop" + 0.005*"repost" + 0.005*"white"
Topic 1: 0.020*"black" + 0.008*"cop" + 0.007*"follow" + 0.006*"u" + 0.006*"people" + 0.006*"america" + 0.005*"like" + 0.004*"american" + 0.004*"police" + 0.004*"life"
Topic 2: 0.014*"u" + 0.013*"police" + 0.013*"black" + 0.007*"officer" + 0.006*"people" + 0.006*"man" + 0.006*"cop" + 0.005*"video" + 0.005*"follow" + 0.005*"like"

Cluster 1 Topics:
Topic 0: 0.027*"black" + 0.018*"police" + 0.010*"bm" + 0.010*"cop" + 0.010*"community" + 0.009*"officer" + 0.009*"matter" + 0.008*"man" + 0.008*"life" + 0.007*"u"
Topic 1: 0.011*"police" + 0.009*"school" + 0.007*"video" + 0.007*"u" + 0.006*"black" + 0.006*"officer" + 0.006*"home" + 0.005*"people" + 0.005*"follow" + 0.004*"dont"
Topic 2: 0.020*"free" + 0.016*"stop" + 0.013*"u" + 0.011*"facemusic" + 0.011*"join" + 0.011*"music" + 0.010*"online" + 0.008*"player" + 0.007*"browser" + 0.007*"ur"

Cluster 2 Topics:
Topic 0: 0.009*"black" + 0.006*"america" + 0.005*"one" + 0.005*"veteran" + 0.005*"american" + 0.004*"usa" + 0.004*"school" + 0.004*"u" + 0.004*"people" + 0.004*"blackmattersuscom"
Topic 1: 0.015*"u" + 0.010*"police" + 0.009*"black" + 0.008*"people" + 0.008*"free" + 0.007*"join" + 0.006*"american" + 0.005*"feel" + 0.005*"bring" + 0.005*"friend"
Topic 2: 0.029*"black" + 0.008*"police" + 0.008*"matter" + 0.007*"people" + 0.007*"follow" + 0.006*"u" + 0.006*"community" + 0.005*"new" + 0.005*"white" + 0.004*"bm"

Cluster 3 Topics:
Topic 0: 0.020*"black" + 0.011*"free" + 0.007*"people" + 0.007*"u" + 0.006*"stop" + 0.005*"music" + 0.005*"community" + 0.005*"one" + 0.005*"like" + 0.005*"join"
Topic 1: 0.011*"black" + 0.009*"police" + 0.008*"people" + 0.008*"u" + 0.007*"join" + 0.005*"state" + 0.005*"school" + 0.005*"like" + 0.005*"student" + 0.005*"right"
Topic 2: 0.024*"black" + 0.014*"police" + 0.009*"matter" + 0.008*"u" + 0.007*"officer" + 0.007*"people" + 0.006*"life" + 0.005*"racism" + 0.004*"community" + 0.004*"blackmattersuscom"

Cluster 4 Topics:
Topic 0: 0.022*"black" + 0.016*"u" + 0.008*"repost" + 0.008*"mexican" + 0.007*"african" + 0.007*"join" + 0.006*"community" + 0.006*"woman" + 0.006*"police" + 0.005*"american"
Topic 1: 0.038*"black" + 0.014*"police" + 0.013*"u" + 0.012*"matter" + 0.012*"america" + 0.011*"join" + 0.008*"community" + 0.007*"people" + 0.006*"racism" + 0.006*"woman"
Topic 2: 0.017*"u" + 0.015*"black" + 0.010*"police" + 0.008*"community" + 0.006*"cop" + 0.006*"stand" + 0.006*"life" + 0.006*"join" + 0.006*"like" + 0.006*"time"
```

Рисунок 3.5 – Тематичні патерни, розподілені по кластерах

Крок 5. Аналіз результатів та прийняття рішень:

– оцінка тематичних патернів та характеристик кожного кластеру;

– визначення, які кластери мають схожі інтереси та споживчі звички, і як це може вплинути на маркетингові стратегії;

– прийняття рішення щодо персоналізованих маркетингових кампаній, спрямованих на кожен кластер.

Крок 6. Експерименти та покращення:

– виконання декількох ітерацій, оптимізуючи параметри методів та аналізуючи результати;

– розгляд можливості додавання інших факторів, таких як демографічні дані, для покращення кластеризації та аналізу.

Ці кроки допоможуть створити персоналізовані маркетингові стратегії для різних груп користувачів у соціальних мережах та забезпечити підтримку прийняття рішень в сфері маркетингу на основі кластеризації [24-26].

3.3 Тестування розроблених застосунків та аналіз результатів

Слідуючи алгоритму реалізації було розроблено застосунок для проведення спектральної кластеризації на даних про рекламу в соціальній мережі Facebook, направлену на американську спільноту, та застосування алгоритму LDA для пошуку тематичних патернів серед текстів представлених рекламних дописів.

Кожен кластер містить наступну кількість точок даних, що представлено на рисунку 3.6.

Для тестування програмного застосунку було вирішено дослідити дані, на відповідність кількості відгуків на допис до дати розміщення допису, розбиті по кожному кластеру. Отримані результати представлені на рисунках 3.7–3.12.

```
Cluster 3.0: 757 data points
Cluster 2.0: 718 data points
Cluster 1.0: 1023 data points
Cluster 4.0: 319 data points
Cluster 0.0: 690 data points
```

Рисунок 3.6 – Кількість точок даних у кожному кластері

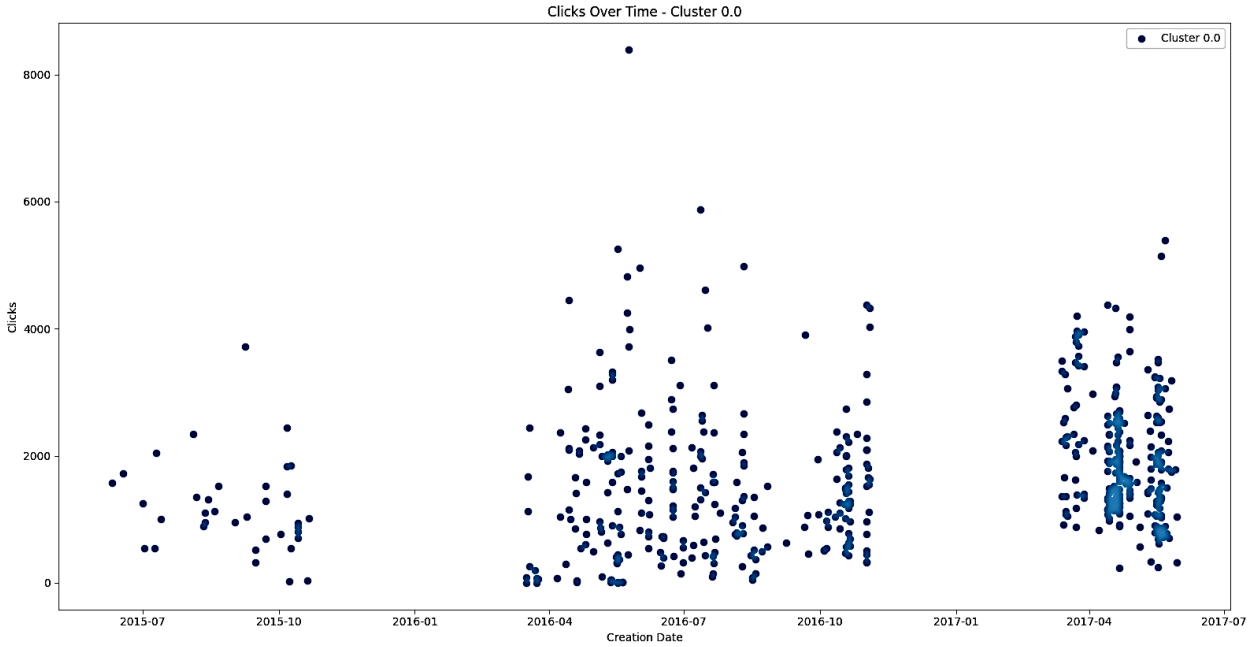

Рисунок 3.7 – Кількість відгуків на допис відносно часу для кластеру 0

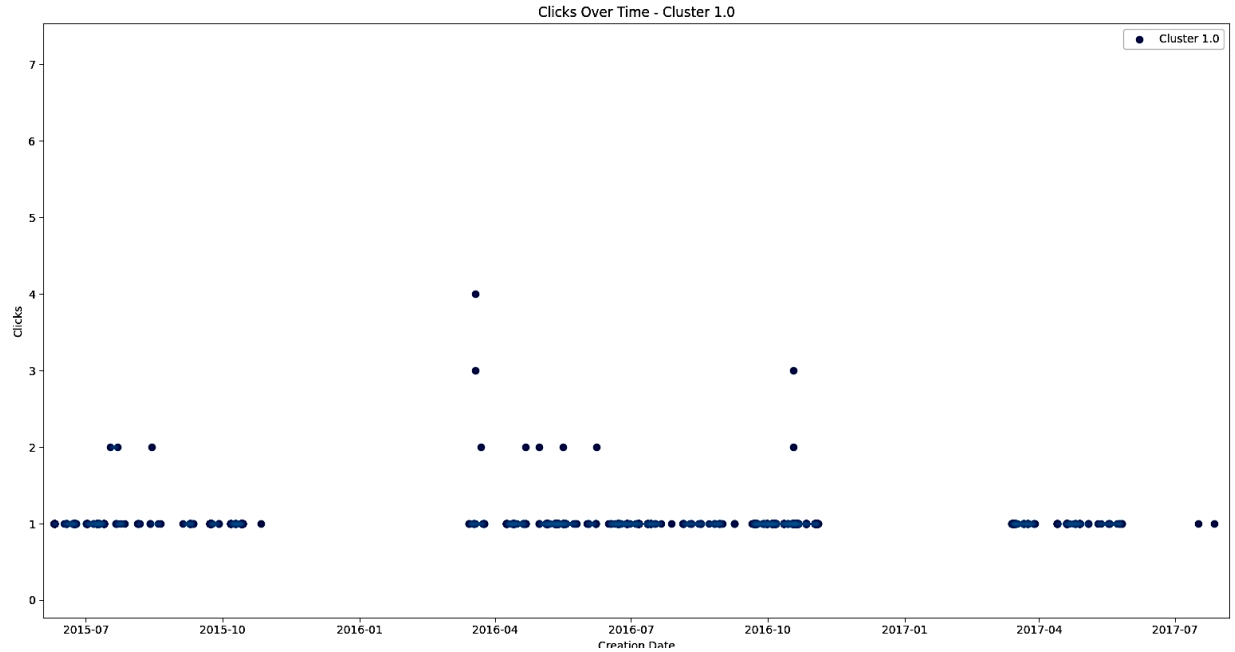

Рисунок 3.8 – Кількість відгуків на допис відносно часу для кластеру 1

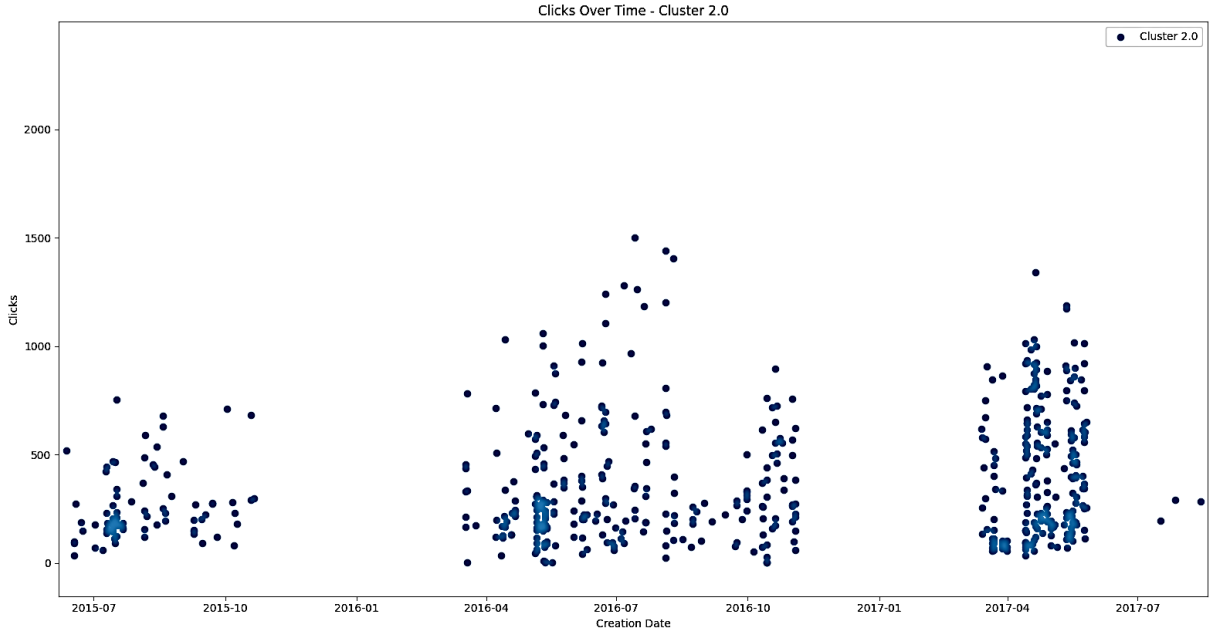

Рисунок 3.9 – Кількість відгуків на допис відносно часу для кластеру 2

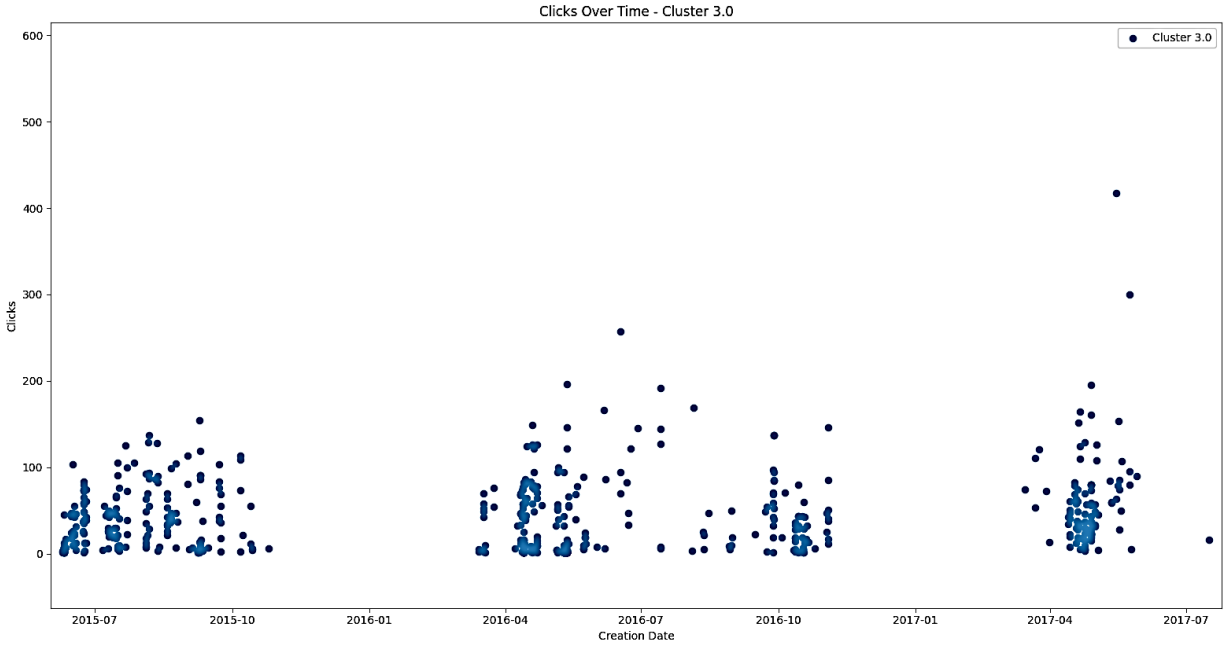

Рисунок 3.10 – Кількість відгуків на допис відносно часу для кластеру 3

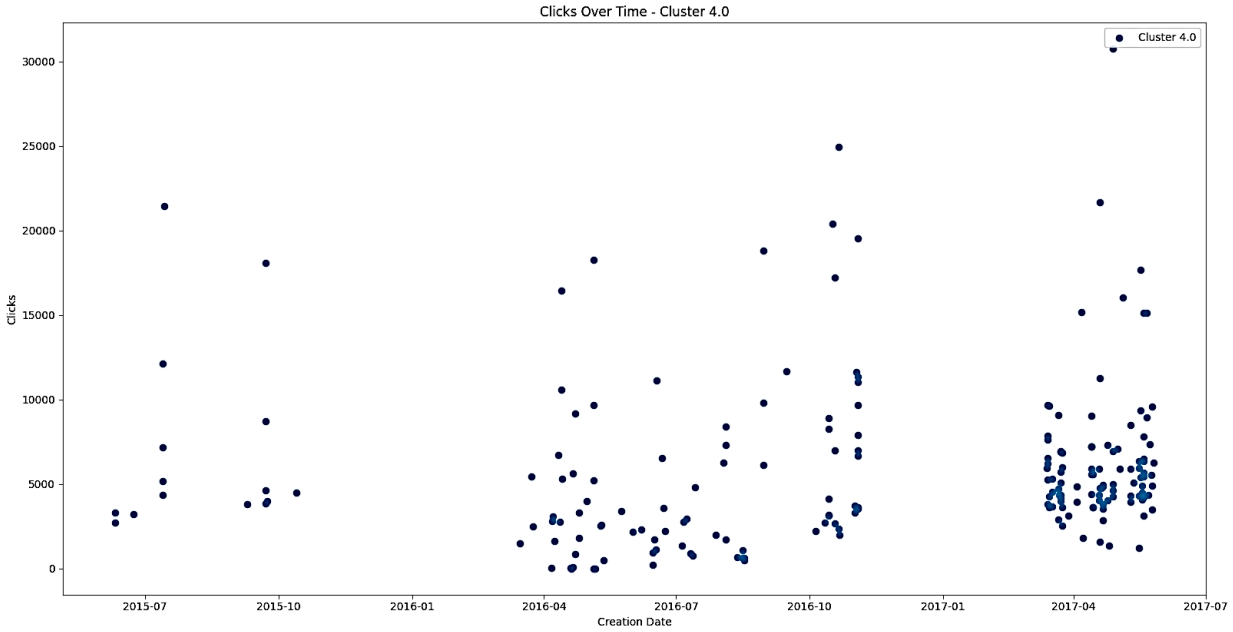

Рисунок 3.11 – Кількість відгуків на допис відносно часу для кластеру 4

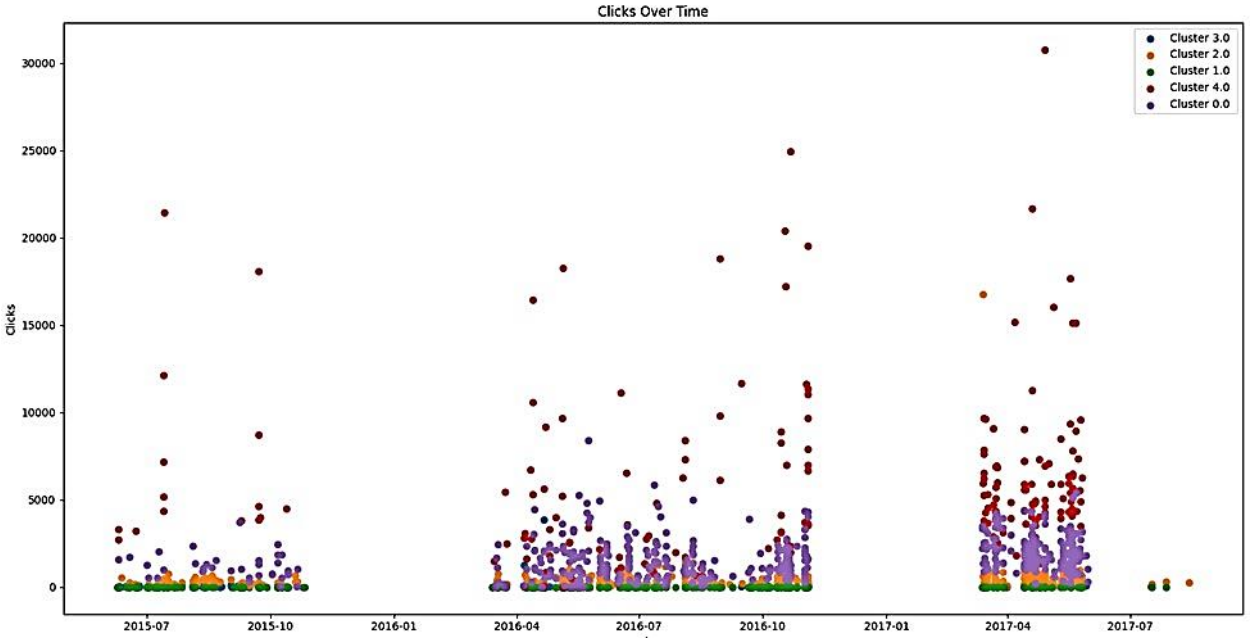

Рисунок 3.12 – Кількість відгуків на допис відносно часу для всіх кластерів

Аналізуючи тематичні патерни для кожного кластеру, можна виділити ключові терміни та описати основні тематичні концепції для обраного набору даних.

Теми кластеру 0:

– тема 0: зосереджена на термінах «black», «people», «police». Обговорення аспектів життя темношкірого населення та стосунків із поліцією;

– тема 1: звертає увагу на «black», «police», «black lives matter», «white». Обговорення питань расової нерівності та актуальних подій;

– тема 2: включає «black», «follow», «america», «police». Обговорюються теми американської спільноти та слідкування за подіями.

Тему кластеру 1:

– тема 0: обговорення «black», «police», «man», «cop», «video». Тема може вказувати на обговорення поліційних втручань та відеозаписів;

– тема 1: включає «join», «black», «police». Тема про активну участь та об’єднання в справі;

– тема 2: містить «free», «black», «stop», «community», «facemusic». Обговорення важливості свободи та темношкірої спільноти.

Теми кластеру 2:

– тема 0: зосереджена на «black», «police», «matter», «life». Пов’язана із питаннями расизму та важливості життя;

– тема 1: включає «police», «black», «people». Про новини та питання, що стосуються поліції;

– тема 2: містить «free», «american», «join», «feel», «friend». Обговорення американської ідентичності та свободи.

Теми кластеру 3:

– тема 0: обговорення «police», «people», «black». Може вказувати на проблеми взаємодії із поліцією та суспільством;

– тема 1: містить «free», «stop», «music». Обговорення важливості свободи музичних тем;

– тема 2: зосереджена на «black», «police», «racism», «people». Може стосуватися питань расової нерівності та життя спільноти.

Теми кластеру 4:

– тема 0: включає «police», «make», «dont». Пов’язана з реакцією на дії поліції та висловлення незгоди;

– тема 1: зосереджена на «black», «join», «matter». Обговорення важливих подій та питань;

– тема 2: містить «black», «community», «like». Може вказувати на обговорення спільноти та подій.

Ці аналізи базуються на частоті ключових термінів та їхніх зв’язках у кожній темі кожного кластеру.

Для демонстрації найбільш часто використовуваних слів в кожній темі використано бібліотеку wordcloud. Результат роботи створеного методу показано на рисунку 3.13.

Рисунок 3.13 – Найбільш уживані слова в темах, виділених алгоритмом LDA

Порівнюючи теми між кластерами, оцінено їхню унікальність та відмінність.

Унікальність: кожен кластер має свої унікальні тематичні аспекти, які відрізняють його від інших. У кластері 0 обговорюються теми, пов’язані із життям темношкірого населення та взаємодією з поліцією, в той час як у кластері 1 акцентується на участі та об’єднанні в поліцейських справах. Це

вказує на те, що кожен кластер має свої власні особливості та тематичні концепції.

Відмінність: є відмінності між темами різних кластерів, що показують різні аспекти обговорень в групах. Наприклад, у кластері 2 тематику зосереджено на питаннях расизму та важливості життя, тоді як у кластері 4 обговорюються реакції на дії поліції та вираження незгоди. Це свідчить про те, що кластери охоплюють різні аспекти тематики [27-29].

Загалом, хоча можуть бути спільні теми, такі як «black», «police», «matter», кожен кластер має свої власні особливості та варіації у фокусі обговорень. Це вказує на різноманітність та широкий спектр тем у спільноті. Для оцінки різноманітності тем, створено теплову карту тематичних патернів, де кожен рядок представляє документ, а кожен стовпець представляє тему. Значення в матриці вказують ступінь належності документа до теми (вагу). (рис. 3.14).

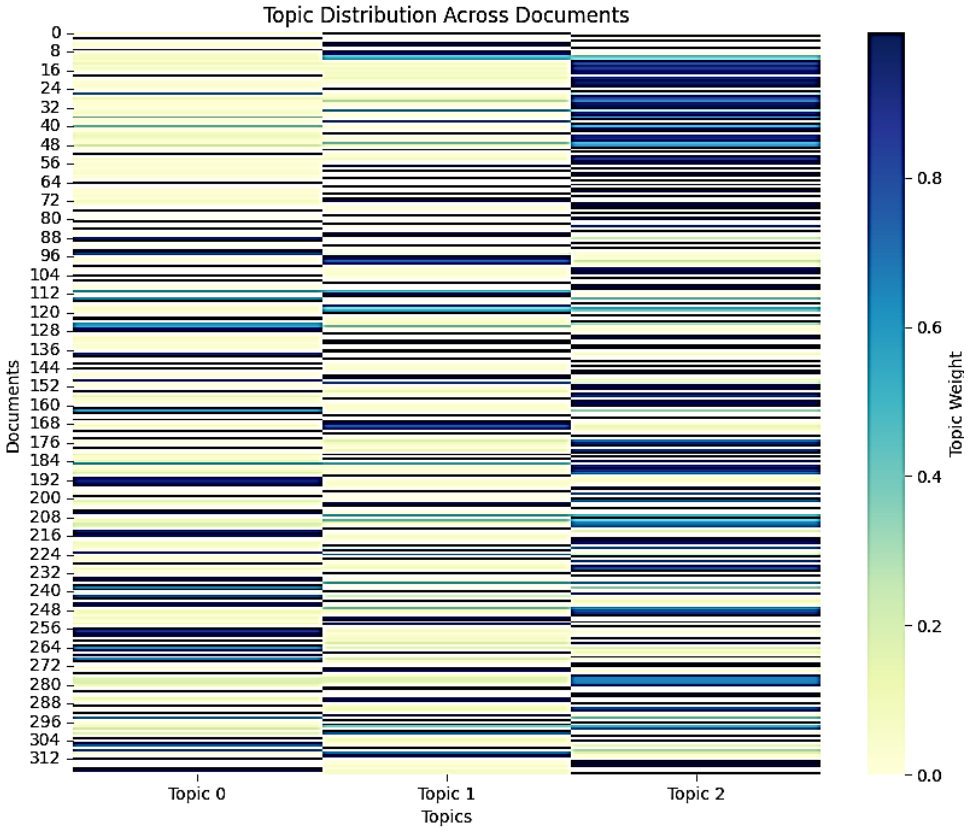

Рисунок 3.14 – Теплова карта тематичних патернів

Аналізуючи результати дослідження, можна сказати, що незалежно від кількості реакцій та відгуків на дописи, теми, що викликають цікавість американської спільноти Facebook, зосереджені на подіях, пов’язаних з рухом Black Lives Matter (міжнародний антирасистський рух активістів), який називає своєю метою боротьбу проти насильства до темношкірого населення. Займається організацією акцій протесту, демонстрацій проти поліційного насильства та расової дискримінації в правовій системі США.

Виходячи з результатів тестування, визначено, що ефективність застосування спектральної кластеризації в поєднанні з алгоритмом LDA є досить високою та показує оптимальний результат для вирішення задач з прийняття рішень [30-33].

3.4 Перспективи подальшої роботи

Робота над кластеризацією у соціальних мережах для підтримки прийняття рішень у сфері маркетингу є динамічним напрямом досліджень, і вона має перспективи для подальшого розвитку.

При детальному розгляді було виявлено наступні можливі напрямки для майбутньої роботи:

– врахування динаміки зміни інтересів: розробка методів, які враховують динаміку зміни інтересів та поведінки користувачів у часі. Маркетингові стратегії повинні адаптуватися до змін у попиті та трендах;

– інтеграція інших джерел даних: врахування додаткових джерел даних, таких як географічні дані, демографічні дані та дані взаємодій користувачів з різними типами контенту, для отримання більш повного зображення користувачів;

– розширення на інші галузі: використання схожих методів для кластеризації користувачів у інших сферах, таких як медицина, освіта або розваги, з метою підтримки прийняття рішень в цих галузях;

– удосконалення методів кластеризації: розвиток нових алгоритмів кластеризації та оптимізація існуючих для кращого виявлення та опису різних груп користувачів. Методи глибинного навчання, автоматичного машинного навчання та зростаючого навчання можуть знайти застосування в цій області;

– аналіз впливу маркетингових заходів: дослідження ефективності персоналізованих маркетингових кампаній, які базуються на кластеризації користувачів, та визначення їх впливу на збільшення залучення та продажів;

– аспекти конфіденційності та етики: розробка та вдосконалення методів для забезпечення конфіденційності та етичної обробки особистих даних користувачів у контексті збільшення важливості цих аспектів;

– застосування обробки природної мови (NLP): використання методів NLP для аналізу текстової інформації, щоб отримати більше інформації про вподобання та схильності користувачів.

Далі, з плином часу та зростанням доступності даних, робота в цьому напрямку може привести до нових інновацій та покращень у розумінні користувачів та оптимізації маркетингових стратегій.

ВИСНОВКИ

У рамках кваліфікаційної роботи було розроблено і реалізовано застосунок з використання методів кластеризації у задачах прийняття рішень (додаток А).

Під час проведення роботи використано алгоритм спектральної кластеризації для групування текстових даних, а також проведено тематичний аналіз отриманих кластерів за допомогою моделі LDA.

Використано бібліотеку scikit-learn для спектральної кластеризації текстових даних. Вхідні дані представлені у форматі CSV, і успішно завантажено та оброблено їх для використання в алгоритмі.

Використано модель LDA для визначення тематичних патернів у кожному з кластерів. Результати LDA показали ключові слова для кожної теми в кожному кластері.

Проведено аналіз тематичних патернів для кожного кластера, визначивши ключові слова та концепції. Виділено унікальні та відмінні аспекти обговорень у кожному кластері.

Кластеризація та аналіз показали різноманіття тем та поглядів в обговореннях. Кожен кластер має свої унікальні особливості та фокус, що свідчить про різноманітність спільноти. Важливо провести більш докладний аналіз та врахувати контекст обговорень для кращого розуміння тематики.

Загальною метою роботи застосунку є розкриття тематичної структури текстових даних, що дозволяє отримати глибше розуміння спільноти соціальної мережі та її реакції на рекламу певних тематик. У подальшому можна розширити аналіз та дослідження конкретних аспектів.

Результати дослідження апробовано у вигляді тез доповідей під час Міжнародної молодіжної конференції «New ways of creating scientific ideas for implementation» [18].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Бойко, Я. В. (2017). Математичне та програмне забезпечення кластеризації ухвалення рішень.

2. Мацуга, О. М., & Ємел'яненко, Т. Г. (2008). Підтримка прийняття рішень під час кластерного аналізу медичних даних. *Актуальні проблеми автоматизації та інформаційних технологій*, (12), 28-36.

3. Baybuz, O. G., & Sidorova, M. G. (2013). Інформаційна технологія кластеризації даних у часовому періоді спостережень. *System research and information technologies*, (4), 59-66.

4. Гнатієнко, Г. М., & Снитюк, В. Є. (2008). Експертні технології прийняття рішень. К.: ТОВ «Маклаут.

5. Струбицька, І. П., & Бойко, Я. В. (2015). Ієрархічний метод кластеризації для задачі ухвалення рішень (Doctoral dissertation, Тернопіль, ТНЕУ).

6. Гороховатський, В. О., & Творошенко, І. С. (2021). Методи інтелектуального аналізу та оброблення даних: навч. посібник.

7. Приставка, О. П., & Сидорова, М. Г. (2011). Підтримка прийняття рішень у задачах кластерного аналізу. *Актуальні проблеми автоматизації та інформаційних технологій*, (15), 115-123.

8. Daradkeh Y.I., Gorokhovatskyi V., Tvoroshenko I., Gadetska S., and Al-Dhaifallah M. (2023) Statistical data analysis models for determining the relevance of structural image descriptions, *IEEE Access*, 11, pp. 126938-126949.

9. Khameneh, A. Z., Kilicman, A., & Ali, F. M. (2022). Transitive fuzzy similarity multigraph-based model for alternative clustering in multi-criteria group decision-making problems. *International Journal of Fuzzy Systems*, 24(5), 2569-2590.

10. Aliahmadipour, L., Eftekhari, M., & Torra, V. (2022). HFC: Data clustering based on hesitant fuzzy decision making. *Iranian Journal of Fuzzy Systems*, 19(5), 167-181.

11. Daradkeh Y.I., Gorokhovatskyi V., Tvoroshenko I., and Zeghid M. (2022) Tools for fast metric data search in structural methods for image classification, *IEEE Access*, 10, pp. 124738-124746.

12. Jia, H., Ding, S., Xu, X., & Nie, R. (2014). The latest research progress on spectral clustering. *Neural Computing and Applications*, 24, 1477-1486.

13. Huang, Z., Zhou, J. T., Peng, X., Zhang, C., Zhu, H., & Lv, J. (2019, August). Multi-view Spectral Clustering Network. In IJCAI (Vol. 2, No. 3, p. 4).

14. Huang, D., Wang, C. D., Wu, J. S., Lai, J. H., & Kwoh, C. K. (2019). Ultra-scalable spectral clustering and ensemble clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6), 1212-1226.

15. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78, 15169-15211.

16. Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys* (CSUR), 54(7), 1-35.

17. Rieger, J., Rahnenführer, J., & Jentsch, C. (2020, June). Improving latent Dirichlet allocation: On reliability of the novel method LDAPrototype. *In International Conference on Applications of Natural Language to Information Systems* (pp. 118-125). Cham: Springer International Publishing.

18. Алевська А. (2023) Дослідження методів кластеризації даних у задачах прийняття рішень, *Abstracts of I International Scientific and Practical Conference «New ways of creating scientific ideas for implementation», (September 18 – 20, 2023). Varna, Bulgaria*, pp. 233-236.

19. M. Ayaz Ahmad, Irina Tvoroshenko, Jalal Hasan Baker, and Vyacheslav Lyashenko (2019) Modeling the Structure of Intellectual Means of Decision-Making Using a System-Oriented NFO Approach, *International Journal of Emerging Trends in Engineering Research*, 7(11), pp. 460-465.

20. Творошенко І.С. (2021). Технології прийняття рішень в інформаційних системах: навч. посібник. *Харків: ХНУРЕ*.

21. Daradkeh Y.I., and Tvoroshenko I. (2020) Technologies for Making Reliable Decisions on a Variety of Effective Factors using Fuzzy Logic, *International Journal of Advanced Computer Science and Applications*, 11(5), pp. 43-50.

22. Gorokhovatskyi V.O., Tvoroshenko I.S., and Vlasenko N.V. (2020) Using fuzzy clustering in structural methods of image classification, *Telecommunications and Radio Engineering*, 79(9), pp. 781-791.

23. Tvoroshenko I. (2020). Information technologies for decision-making on the conditions of spatially distributed objects, in *Abstracts of I International Scientific and Practical Conference. Problems and perspectives of modern science and practice*, Austria. pp. 45-50.

24. Pomazan V., Tvoroshenko I., and Gorokhovatskyi V. (2023) Development of an application for recognizing emotions using convolutional neural networks, *International Journal of Academic Information Systems Research*, 7(7), pp. 25-36.

25. Pomazan V., Tvoroshenko I., and Gorokhovatskyi V. (2023) Handwritten character recognition models based on convolutional neural networks, *International Journal of Academic Engineering Research*, 7(9), pp. 64-72.

26. Tvoroshenko I., Gorokhovatskyi V., Kobylin O., and Tvoroshenko A. (2023) Application of deep learning methods for recognizing and classifying culinary dishes in images, *International Journal of Academic and Applied Research*, 7(9), pp. 57-70.

27. Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. California management review, 61(4), 66-83.

28. Gorokhovatskyi V., Tvoroshenko I., Kobylin O., and Vlasenko N. (2023) Search for visual objects by request in the form of a cluster representation for the structural image description, *Advances in Electrical and Electronic Engineering*, 21(1), pp. 19-27.

29. Gati, I., Levin, N., & Landman-Tal, S. (2019). Decision-making models and career guidance. International handbook of career guidance, 115-145.

30. Гороховатський В.О., Творошенко І.С., Чмутов Ю.В. (2022) Застосування систем ортогональних функцій для формування простору ознак у методах класифікації зображень, *Сучасні інформаційні системи*, 6(3), С. 5-12.

31. Гороховатський В., Передрій О., Творошенко І., Марков Т. (2023) Матриця відстаней для множини компонентів структурного опису як інструмент для створення класифікатора зображень, *Сучасні інформаційні системи*, 7(1), С. 5-13.

32. Гороховатський В., Творошенко І., Сидоренко Д. (2021) Класифікація зображень із використанням кластерного подання, *Міжн. наук. симпозіум Інтелектуальні рішення-С. Обчислювальний інтелект. Теорія прийняття рішень: праці міжн. наук. симп. (Вересень 29, 2021)*. Київ-Ужгород, С. 44-45.

33. Gorokhovatskyi V., Tvoroshenko I. (2023) Identification of visual objects by the search request. *International scientific symposium «INTELLIGENT SOLUTIONS-S». Computational intelligence (results, problems and perspectives). Decision making theory: proceedings of the international symposium*, September 28, 2023, Kyiv-Uzhorod, Ukraine, pp. 25-27.