Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____*Комп'ютерних наук*_____
(повна назва)

Кафедра _____*Системотехніки*_____
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____*другий (магістерський)*_____

_____*Дослідження методів кластерного аналізу для визначення*_____
_____*функціонального та психічного стану користувачів соціальних мереж*_____
(тема)

Виконав:
студент _*2*_ курсу, групи _____*СПРм-19-2*____
_____*Янченко В.В.*_____
(прізвище, ініціали)

Спеціальність _*122 — Комп'ютерні науки*_____
(код і повна назва спеціальності)
Тип програми _____*освітньо-наукова*_____
(освітньо-професійна або освітньо-наукова)
Освітня програма __*Системне проектування*___
(повна назва освітньої програми)

Керівник _____*доц. Тітов С.В.*_____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ ____*Гребеннік І.В.*_____
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет_____*Комп'ютерних наук*_____

Кафедра_____*Системотехніки*_____

Рівень вищої освіти_____*другий (магістерський)*_____

Спеціальність_____*122 — Комп'ютерні науки*_____
(код і повна назва)

Тип програми _____*освітньо-наукова*_____
(освітньо-професійна або освітньо-наукова)

Освітня програма_____*Системне проектування*_____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

«____»_____ 20 ___ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові_____*Янченко Вадиму Вячеславовичу*_____
(прізвище, ім'я, по батькові)

1. Тема роботи _____*Дослідження методів кластерного аналізу для визначення*_____
_____*функціонального та психічного стану користувачів соціальних мереж*__
затверджена наказом університету від __*19*__ ____*04*_____*2021 р.* № __*509 Ст*__

2. Термін подання студентом роботи до екзаменаційної комісії _*24*___ ___*05*_____ *2021 р.*

3. Вихідні дані до роботи _*користувач додатку, дані користувачів мережі Twitter, методи кластерного аналізу, методи аналізу тональності, пояснювальна записка*

4. Перелік питань, що потрібно опрацювати в роботі *вступ, мета роботи, аналіз предметної галузі і постановка задачі, аналіз можливих методів кластеризації, приведення текстових даних до формату який можливо кластеризувати, аналіз тональності тексту, аналіз результатів кластеризації* _____
_____
_____

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) *Плакати на аркушах формату А4: схема ієрархії критеріїв порівняння методів кластерного аналізу, схема структури ієрархії категорій вдоволеності, схема глобальних пріоритетів, алгоритм "k-means", демонстрація алгоритму "k-means", алгоритм "c-means", порівняння методів "k-means" та "c-means", алгоритм ієрархічної кластеризації, приклад ієрархічної кластеризації областей України за середньою заробітньою платнею, схема аналізу даних на мові програмування R, основне меню додатку, вікно додавання даних користувача, форма помилки, вікно з описом метрик психічного стану користувача, результати аналізу тональності даних користувачів, візуалізація дендрограми ієрархічної кластеризації*

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
Спецчастина	доц. Тітов С.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	*Отримання, аналіз завдання, уточнення*	*23 березня 2021 р.*	
2	*Виявлення пріоритетних властивостей методів кластерного аналізу*	*27 березня 2021 р.*	
3	*Постановка задачі та вибір методу її вирішення*	*03 квітня 2021 р.*	
4	*Проведення експериментальних досліджень*	*20 березня 2021 р.*	
5	*Оформлення пояснювальної записки*	*25 квітня 2021 р.*	
6	*Підготовка презентації та доповіді*	*05 травня 2021 р.*	
7	*Подання закінченої роботи науковому керівнику*	*13 травня 2021 р.*	
8	*Подання роботи на рецензування*	*14 травня 2021 р.*	
9	*Попередній захист*	*16 травня 2021 р.*	
10	*Подання роботи до комісії*	*24 травня 2021 р.*	

Дата видачі завдання __*29*___ __*03*__ *2021* р.

Студент _____ *Янченко В.В*
(підпис)

Керівник роботи _____ *доц. Тітов С.В.*
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи магістра, 72 с., 16 рисунків, 11 таблиць, 20 джерел.

КВАЛІФІКАЦІЙНА РОБОТА, АНАЛІЗ ДАНИХ, МЕТОД K-MEANS, TWEETER, AFINN-111, BIG DATA, КЛАСТЕРНИЙ АНАЛІЗ, ІЄРАРХІЧНИЙ АНАЛІЗ, АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТУ.

Об’єкт дослідження — програмні додатки, що використовують соціальні мережі для отримання даних про їх користувачів та алгоритми кластерного аналізу та аналізу тональності тексту для обробки вхідних даних.

Предмет дослідження — алгоритми аналізу тональності тексту, що базується на словнику тональності та алгоритми кластерного аналізу для обробки семантично-оброблених даних для визначення груп користувачів за психічним та функціональним станом.

Мета кваліфікаційної роботи — дослідження методів кластерного аналізу у рамках задачі аналізу тональності текстових даних користувачів соцмереж.

Методи досліджень — метод аналізу ієрархій, метод попарних порівнянь, метод експертних оцінок.

Результат кваліфікаційної роботи — знайдений найбільш ефективний метод кластерного аналізу для задачі кластеризації даних аналізу тональності тексту, розроблений додаток на мові програмування C# для визначення психічного та функціонального стану користувачів соціальних мереж, розроблений додаток на мові програмування R для кластерного аналізу даних.

Галузь застосування — оптимізація процесу найму робітників на державні/цивільні посади, підвищення ефективності психо/медичної терапії шляхом моніторингу настроїв пацієнта, попередження соціальних загроз (терактів, соціальних збурень і т.п.), таргетований аналіз настроїв груп людей у рамках маркетингових кампаній.

ABSTRACT

Master’s explanatory note: 72 pages, 16 figures, 11 tables, 20 sources.

MASTER’S THESIS, DATA ANALYSIS, K-MEANS, TWEETER, AFINN-111, BIG DATA, CLUSTER ANALYSIS, HIERARCHICAL ANALYSIS, SENTIMENT ANALYSIS.

The object of research — monitoring systems that use social networks to obtain users data, both cluster analysis and sentiment analysis algorithms used to process input data.

The subject of research — text sentiment analysis algorithms, based on the sentiment dictionary and cluster analysis algorithms for processing semantically preprocessed data to determine user groups by mental and functional state.

The purpose of the qualification work — to study the methods of cluster analysis within the problem of social network users data sentiment analysis.

Research methods — system analysis, pairwise comparison, methods of text analysis based on sentiment dictionaries, methods of cluster analysis.

The result of the qualification work — the most efficient method of cluster analysis is found for the task of clustering the data prepared by sentiment analysis, developed an application with C# programming language to determine the mental and functional state of social network users, developed an application in R programming language for sentiments data cluster analysis.

Application domains — optimization of hiring process of workers for governmental/civil positions, improving the effectiveness of psycho/medical therapy by monitoring the patient's psychological state, prevention of social threats (terrorist attacks, social disruptions, etc.), targeted analysis of the mood of groups of people in marketing campaigns.

ЗМІСТ

ВСТУП

Наприкінці 50-х років XX сторіччя протягом “холодної війни” Міністерством оборони США була створена організація, яка займалась питанням можливості з’єднання комп’ютерів, що були розташовані у різних місцях. Створену технологію, що надавала можливість об’єднувати комп’ютери у мережі та обмінюватися даними на досить, на той час, високій швидкості, пізніше назвали Інтернет. З плином часу технологія привернула увагу деяких навчальних закладів та окремих науковців, а згодом її перевагами почнуть потроху користуватися й звичайні громадяни. На початку XXI століття використання Інтернету стало стрімко поширюватися тому, що у всі сфери діяльності людини все більш активно стали вторгатися персональні комп’ютери, а з появою смартфонів та технології 4G доступ до Інтернету з’явився майже у кожної людини цілодобово.

Користувачі мережі Інтернет, мають широкі можливості для отримання різноманітної соціальної, наукової, технологічної та іншої поточної інформації. Інтернет являє доступ до величезної безлічі джерел електронної інформації. Сотні тисяч електронних каталогів, баз даних, архівів технічної і програмної документації, бібліотек програм, науково-технічних довідників, електронних газет і журналів, бюлетенів новин і багатьох інших інформаційних матеріалів можна отримати по каналах цієї глобальної міжнародної мережі безпосередньо на робоче місце.

Інтернет значно впливає на сучасне інформаційне суспільство. Він вносить значні зміни у всі сфери його життя: навчання, роботу, побут, дозвілля.

Феномен Інтернету став справжнім технологічним проривом і відкрив незкінченну кількість можливостей для звичайних людей. Зараз без Інтернету неможливо уявити наше суспільство. Згідно з дослідженнями компанії IBM щоденно люди генерують близько двох з половиною мільйонів терабайт даних.

За допомогою Інтернету людство вирішує важливі соціальні сфери людини, такі як обговорення екологічних проблем, можливість навчання у різних місцях планети та швидке розповсюдження новітніх способів лікування. Простота використання інтернет-технологій велику кількість покупців і продавців товарами, послугами, інформацією, в свою чергу відносини між суб'єктами економічної діяльності сформували новітні моделі ведення бізнесу. З високою часткою вірогідності можна стверджувати, що відмова від Інтернету у найближче століття неможлива через його цілковите розповсюдження у більшість сфер людського життя. Інтернет-технології легко дозволяють спілкуватися з величезною кількістю людей, дотримуючись при цьому відносної анонімності, швидко та ефективно долаючи державні границі. Однак не слід забувати, що ті ж технології, що сприяють такому спілкуванню, можуть використовуватися в поганих цілях. Розповсюдження Інтернету породжує і масу новітніх проблем. Даркнет, подібно до темної сторони Місяця, десятки років бентежить розуми людей, котрі все більше і більше сходяться у думці, що дана технологія містить певні загрози суспільству. Також небезпеку суспільству можуть становити і будь-які інші люди, що використовують Інтернет з лихими намірами, такі як терористи чи хакери.

Тому у 2011 році на Міжнародному симпозіумі з інформаційної безпеки академік Фан Биньсин вперше висунув концепцію «суверенного інтернету». На території Китайської народної республіки доступ до багатьох закордонних сайтів заблокований завдяки системі, яка фільтрує контент у Інтернеті. В основному ідеї кіберсуверінетету КНР складаються з чотирьох принципів:

– кожна країна повинна володіти повним контролем над своїм сегментом Інтернету;

– держава повинна мати можливість захищати свій сегмент Інтернету від будь-яких зовнішніх атак;

– всі країни повинні мати рівні права на використання ресурсів Інтернету;

– інші країни не повинні мати можливість контролювати корневі DNS-сервера, через які здійснюється доступ до національного сегменту Інтернета.

Існує думка, що Інтернет небувалою популярність завдячує можливості вільного висловлювання власних думок, якою користуються багато людей по всьому світу, що побачили відлуння лібертаріанських ідей у цій технології. Соціальні платформи накшталт Reddit та Twitter стали центрами гарячих диспутів стосовно тем, що давно бентежили людство, і стали підтвердженням вислову “істина народжується у суперечці”. Проте бізнес-моделі Google, Amazon і Facebook засновані на використанні Інтернет-технологій для збору та монетизації персональних даних. Тому різниця між лібертаріанським поглядом та обмеженнями в «китайському стилі» розмивається. На перший план виходить проблема конфіденційності – право власності на персональні дані.

Криза Covid-19 дозволила поглянути на проблему стеження та конфіденційності з нового боку. І тут Китай та США демонструють нам два різних способи вирішення проблеми. На перші ознаки нових спалахів захворювання Китай реагує суворим карантином, обов’язковим тестуванням, ношенням масок та використанням системи відстеження контактів за допомогою мобільних додатків та QR-кодів. В США вважають такі заходи неприйнятним порушенням у вільному та відкритому суспільстві.

Незважаючи на те, що в наш час не існує єдиного погляду на проблему стеження держав за власними громадянами, і увесь час точаться обговорення щодо доцільності таких дій. Здебільшого, цивілізовані країни, що поважають права та свободи своїх громадян, переслідують лише добрі наміри, такі як попередження різноманітних загроз чи покращення рівня життя населення.

Інструментом аналізу стану людини, з метою недопущення загроз може стати аналіз тональності, що у зв’язку з кластерним аналізом дозволить обробляти (кластеризувати) значні об’єми інформації. При додатковому використанні таких оброблених даних для навчання нейронних мереж можна навіть передбачити потенційні загрози від користувачів соціальних мереж.

1 СИСТЕМНИЙ АНАЛІЗ ПРОБЛЕМИ

1.1 Аналіз предметної галузі

Мережа інтернет, а зокрема і соціальні мережі, розвиваються раніше небаченими темпами. З кожним днем кількість документів, що висловлюють певні думки, містять коментарі, відгуки, критику, огляди тощо збільшується у геометричній прогресії. Такі дані містять цінну інформацію, яка може допомогти людям у прийнятті певних рішень у різних галузях. Наприклад, огляди товарів можуть допомогти підприємствам просувати свою продукцію; коментарі щодо політичних подій можуть допомогти політикам зкорегувати свою політичну стратегію; навіть критика може допомогти сторонам суперечки задуматися над правильністю своєї думки та позиції. Однак кількість таких документів величезна, тому людям неймовірно складно власними силами прочитати та проаналізувати їх усі. Таким чином, автоматичний аналіз думок, висловлених на різних веб-платформах, стає все більш важливим для ефективного прийняття рішень. Одним з підходів до аналізу думок є аналіз тональності або аналіз думок. Даний підхід відноситься до широкої області обробки природної мови, обчислювальної лінгвістики та інтелектуальної обробки тексту.

Завдання такого аналізу є досить складним та потребує використання символічних статистичних методів, оскільки онлайн-огляди часто є неструктурованими, суб’єктивними та важкими для опрацювання за короткий проміжок часу. Люди висловлюють свої думки по-різному, тому аналіз тональності у реченні інколи важко здійснити за допомогою відомих статистичних підходів, так званих “безконтрольних” підходах, заснованих лише на лексиконі. З іншого боку, аналіз тональності дуже чутливий до специфіки предметної галузі у випадку застосування контрольованого навчання нейронних мереж. У багатьох доменах відсутня достатня кількість маркованих

даних, без яких неможливе застосування алгоритмів контрольованого машинного навчання. Як результат, при застосуванні розробленого класифікатора на даних з нового домену, які не промарковані у достатньому обсязі, класифікатор тональності показує досить низьку точність.

Аналіз настроїв, що базується на методах кластерного аналізу, — це перспективний підхід до аналізу думок, висловлених у відгуках, коментарях або на форумах. На відміну від двох традиційних підходів до аналізу настроїв, а саме контрольованого навчання та безконтрольних методів заснованих на лексиконі, підхід, який базується на кластеризації, здатний досягти високої точності аналізу, не потребує участі людини у процесі роботи, відносно швидкий та ресурсо-ефективний.

В даній галузі існує два основних напрями дослідження. Перший полягає у класифікації великої кількості думок за принципом біполярних орієнтацій, тобто думки можуть бути позитивні чи негативні. Другий напрямок дослідження полягає у визначенні того, наскільки думки є суб'єктивними чи об'єктивними. Тобто головна задача методик даного напряму — віднесення тексту до одного з двох класів: об’єктивного чи суб’єктивного.

Ці два напрямки досліджень взаємодоповнюють один одного, оскільки при класифікації заснованій на полярності визначення об’єктивних думок у тексті є досить логічним доповнюючим етапом. І навпаки, слова чи фрази, що виражають певну думку, можуть бути ознакою суб’єктивості тексту. Розвиток даних напрямів допоможе людям отримати більше цінної інформації з різних видів тексту.

Дослідження, що проводиться у даній кваліфікаційній роботі, базується на класифікації тексту повідомлень користувачів соціальних мереж за полярністю та охоплює як аналіз окремих слів тексту так і цілих повідомлень та груп повідомлень.

Оскільки як об’єми даних так і число користувачів соціальних мереж збільшується щорічно, виникає потреба у відтворенні дієвих алгоритмів для

перетворення значної кількості даних у корисну інформацію, іншими словами кластеризації, що у свою чергу вказує на актуальність даного напряму дослідження. На основі аналізу тональності даних користувачів соціальних мереж, є можливість стверджувати про їх функціональний та психічний стан. Це досягається засобами машинного навчання, що дають змогу проводити аналіз та робити висновки на основі кластеризованих даних. Це, до прикладу, може потенціально надавати можливість вчасно приймати необхідні рішення, спираючись на рівень суспільної загрози, що може потенційно надходити від окремих користувачів соціальних мереж.

1.2 Огляд і аналіз існуючих методів і засобів вирішення проблеми

1.2.1 Огляд методів аналізу тональності

Традиційним методом аналізу текстів є визначення груп слів, що мають позитивний емоційний відтінок, негативний відтінок або є нейтральними. Цей метод зазвичай називають аналізом тональності або сентимент-аналізом, що є видом текстового аналізу. Головна мета цього методу — визначення емоційної оцінки думки автора про об’єкт чи тему загалом, яка згадується у певному тексті даного автора.

Існує два основних підходи до аналізу тональності:

- контрольоване машинне навчання або методи глибокого навчання;
- безконтрольні підходи, засновані на лексиконі;

Для першого підходу нам зазвичай потрібні попередньо позначені, або “промарковані” дані. Другий підхід покладається на словник слів, попередньо оцінених спеціальним алгоритмом машинного навчання.

Не дивлячись на те, що методи глибокого навчання досягають гарних результатів у точності оцінки тональності тексту, вони дуже ресурсоємкі. Вони вимагають велику кількість даних для тренування що мають бути заздалегідь промарковані, тобто віднесені до певних класів. Процес навчання мережі також досить затратний по відношенню до часу його виконання. Також Контрольоване машинне навчання дуже залежить від предметної області даних, на яких проводиться тренування. До прикладу, досить важко досягти високої точності у аналізі тональності даним методом якщо використовувати дані на основі оглядів фільмів, тому що навіть у позитивних оглядах фільмів досить часто зустрічаються згадки неприємних сцен кіно, в той час як у негативних оглядах часто зустрічаються згадки певних приємних сцен фільму, що все ж сподобались глядачу не дивлячись на загальну негативну оцінку. Іншими словами, даному методу бракує узагальненості.

Типовий підхід до аналізу настроїв у безконтрольних методах полягає в тому, щоб побачити, скільки слів у тексті мають входження у заздалегідь визначеному списку слів, пов'язаних із настроєм. Фраза “у мене поганий день”, може набрати “1” за шкалою негативних настроїв за наявність “поганий” або 0,17, якщо враховувати, що одне з шести слів є негативним. Деякі системи настроїв класифікують слова за шкалою, таким чином що "приголомшливий" може мати оцінку 5, тоді як "чудовий" оцінку 1.

Деякі системи виходять за рамки лише чітко позитивних або негативних забарвлень. Наприклад, програма “LIWC” розширює ідею аналізу настроїв з метою вимірювання десятків інших атрибутів слів, таких як “тон”, “аналітичне мислення” та “соціальний вплив”.

Також існують бібліотеки що використовують алгоритми машинного навчання мовної обробки, також відомі як “NLP”, до прикладу TextBlob. Даний семантичний словник оцінює слова як за полярністю (проте на відміну від AFINN оцінка полярності нормована і лежить у межах [-1;1]) так і за суб’єктивністю. В даному випадку оцінка суб’єктивності знаходиться у межах

[0;1], де 0 означає те, що слово є дуже об’єктивним, а 1 — занадто суб’єктивним.

Тож, загалом, ці методи можна використовувати, коли для дослідження існує набір даних у вигляді слів і є необхідність знайти кількість їх появлення у наборі текстів. Такі методи зазвичай називають «словниковими методами».

В залежності від задачі дослідження, аналіз тональності може мати наступні варіації:

– аналіз документів — визначення яку тональність, позитивну чи негативну має документ вцілому;

– аналіз речень — визначення того чи являється речення позитивно-, негативно- або нейтрально-забарвленим;

– аналіз аспектів — визначення чи є частина речення певною категорією (їжа, спорт тощо) та визначення її полярності.

Слід також зазначити, що слова — це не єдине представлення людських емоцій. Користувачі соціальних мереж також можуть використовувати різні смайли чи мультимедіа компоненти. Для аналізу настроїв таких даних існують окремі програмні бібліотеки. Проте принцип оцінки та шкала дуже схожі на підхід AFINN.

Загалом процес знаходження оцінки тональності тексту залежить від обраного підходу, для безконтрольних підходів, заснованих на лексиконі, алгоритм аналізу тональності можна описати наступними кроками:

– парсинг та підготовка тексту, за необхідності перетворення у зручний формат;

– підготовка словника тональності — списку позитивних і негативних полярних слів з деякою оцінкою, пов’язаною з ними;

– присвоєння оцінки тексту використовуючи різні технічні прийоми, такі як положення слова у рядку, сусідні слова, контекст, частини мови, фрази;

– отримання остаточної оцінки аналізу тональності після агрегації балів окремих частин тексту.

На даний момент існує велика кількість різних словників тональності, що можуть бути застосовані в залежності від потреб дослідження:

- словник AFINN;
- словник MPQA subjectivity;
- словник SentiWordNet;
- словник VADER;
- словник TextBlob.

Застосування аналізу настроїв дуже різноманітні і можуть включати:

- автоматизований аналіз настроїв електронної пошти;
- автоматизований аналіз настроїв коментарів;
- оптимізація порад пошукових запитів (“SEO optimization advice”);
- аналіз настроїв на основі онлайн статей в Інтернеті;
- автоматичне оцінювання настроїв.

До прикладу, завдяки аналізу тональності з метою моніторингу стану бренду можна отримати повне уявлення про те, яке враження певний бренд, товар чи компанія справляє на клієнтів або акціонерів. Публічно доступні засоби масової інформації, такі як відгуки про товари на форумах чи пости в соціальних мережах, можуть розкрити ключову інформацію про те, що бізнес робить правильно чи неправильно. Аналіз настроїв можна використати щоб оцінити вплив нового продукту, рекламної кампанії чи реакції споживача на останні новини компанії в соціальних мережах.

Аналізу тональності у сфері обслуговування клієнтів допомагає працівникам відділів обслуговування клієнтів автоматично сортувати вхідну електронну пошту користувачів на "термінові" або "не термінові" сегменти на основі настроїв електронних повідомлень, проактивно виявляючи розчарованих користувачів. Потім першочергово допомога надається тим користувачам, які мають найбільш термінові потреби.

Використовуючи аналіз тональності для дослідження та аналізу ринку, можна зрозуміти, чому споживачі реагують або не реагують на певні події. Підхід може бути використаний у галузях політології, соціології та психології для аналізу тенденцій, ідеологічних упереджень, думок, оцінки реакцій тощо.

1.2.2 Огляд методів кластерного аналізу

Кластерний аналіз — це процес групування набору об’єктів таким чином, щоб об’єкти однієї групи (яка називаються кластерами) були більш схожі (за певними ознаками) один на одного, ніж об’єкти у інших групах. Він є головним завданням дослідницького аналізу даних і використовується як техніка статистичного аналізу даних, яка використовується насамперед у багатьох галузях, пов’язаних з комп’ютерними науками, включаючи:

- розпізнавання образів;
- аналіз зображення;
- пошук інформації;
- біоінформатика;
- компресія даних;
- комп’ютерна графіка;
- машинне навчання.

Кластерний аналіз сам по собі є не одним конкретним алгоритмом, а виступає загальним завданням, яке потрібно вирішити. Для аналізу даних можна використовувати різні алгоритми, які, здебільшого, суттєво різняться між собою у двох поняттях: визначенні того, що являє собою кластер і як ефективно отримати кластер. До розповсюджених визначень поняття “кластер” можна віднести: група об’єктів з невеликими відстанями між членами групи,

група об’єктів з високою щільністю розміщення, інтервал або певний статистичний розподіл. Таким чином, кластерний аналіз можна охарактеризувати як багатоцільову задачу оптимізації. Вибір алгоритму та параметрів кластеризації (включаючи такі параметри, як функція знаходження відстані між об’єктами кластеру, гранична межа щільності розміщення об’єктів або кількість очікуваних кластерів) залежить від конкретного набору даних (особливостей даних, предметної галузі) та того, яким чином планується використовувати результати кластеризації (маркування даних, знаходження аномалій наборів даних тощо). За своєю природою кластерний аналіз не є автоматичним завданням, а містить у основі ітераційний процес виявлення корисної інформації (знань) — інтерактивну багатоцільову оптимізацію, що включає в себе випробування та невдачі. У процесі аналізу часто доводиться модифікувати алгоритм попередньої обробки даних та параметри моделі, до тих пір, доки результат аналізу не досягне бажаних показників точності.

Кластерний аналіз був заснований у галузі антропології докторами Драйвером і Кребером в 1932 р., а пізніше адаптований у галузі психології Джозефом Зубіном у 1938 р. та Робертом Тріоном у 1939 р. і відомий використанням Кеттеллом у 1943 р. для класифікації теорії людських рис в психології особистості.

Дотепер поняття "кластер" не було визначене однозначно, що є однією з причин того, чому існує так багато алгоритмів кластерного аналізу. Проте все ж науковці знаходять один спільний знаменник: кластер — це група об’єктів даних. Однак різні дослідники використовують різні кластерні моделі, і кожна з цих кластерних моделей визначає свій набір алгоритмів кластерного аналізу. Поняття кластера, суттєво відрізняється за своїми властивостями від алгоритма до алгоритму. Розуміння "кластерних моделей" є ключовим етапом для розуміння відмінностей між різними алгоритмами кластерного аналізу. Типові моделі кластеризації включають:

– моделі зв’язку: наприклад, ієрархічна кластеризація будує моделі на основі зв’язку відстаней між об’єктами;

– моделі центроїдів: наприклад, алгоритм k-середніх, що представляє кожен кластер одним вектором середнього значення;

– моделі розподілу: кластери моделюються з використанням статистичних розподілів, таких як багатовимірний нормальний розподіл, що використовується алгоритмом максимізації очікувань;

– моделі щільності: наприклад, алгоритми DBSCAN і OPTICS визначають кластери як пов'язані в просторі даних області з високою щільністю;

– моделі на основі графів: до прикладу, кліка, тобто підмножина вузлів у графіку, у якій кожні два вузли в підмножині з'єднані ребром, що може розглядатися як прототипова форма кластера;

– нейронні моделі: найвідомішою нейронною мережею з неконтрольованим навчанням є самоорганізуючий словник, такі моделі, як правило, можуть бути подібними до вищезазначених моделей.

Кластеризація — це, по суті, набір таких кластерів, що зазвичай міститять усі об'єкти з набору даних. Крім того, вона може визначати взаємозв'язки кластерів між собою, наприклад, ієрархію кластерів, вбудованих один в одного. Кластеризацію можна класифікувати наступним чином:

– чітка кластеризація: визначає, що кожен об’єкт однозначно або належить певному кластеру або ні;

– нечітка кластеризація (або м’яка кластеризація): кожен об'єкт певною мірою належить кожному кластеру (існує ймовірність належності до об’єкта до кожного кластера).

Як згадано вище, алгоритми кластеризації можна класифікувати на основі їх кластерної моделі і серед них не існує об’єктивно "правильного" алгоритму кластеризації. Здебільшого, якщо не існує математично-

обгрунтованої причини віддати перевагу конкретній кластерній моделі у рамках певної задачі, найбільш дієвий алгоритм кластеризації для конкретної проблеми обирається експериментально.

В останні роки докладено значних зусиль для підвищення ефективності існуючих алгоритмів. Серед нових алгоритмів можна відзначити CLARANS та BIRCH. З плином часу у зв'язку з шаленими темпами росту об'ємів цифрових даних все більш актуальною стає проблема обробки таких великих обсягів інформації (одна з задач напрямку Big Data), що у свою чергу провокує зміну пріоритету з важливості семантичної значущості кластерів, отриманих у результаті аналізу, на важливість швидкості обробки даних. Це призводить до розробки методів попередньої кластеризації, таких як "canopy clustering", що в змозі ефективно обробляти величезні масиви даних, та надавати кластери, що характеризуються досить нечітким групуванням за ознаками та відносно низькою точністю, проте слугують потужним базисом для подальшого аналізу з метою збільшення точності кластеризації за допомогою існуючих більш повільних класичних методів, таких як k-means.

При аналізі багатовимірних даних більшість існуючих алгоритмів неефективні через проблему багатовимірності, що полягає у складності використання більшості функцій знаходження відстані для випадку багатовимірного простору. Це призвело до появи нових алгоритмів кластерного аналізу для багатовимірних даних, які фокусуються на кластеризації підпростору (коли для кластеризації використовуються лише підмножина ознак) та кореляційній кластеризації, що базується на корельований підпростір кластерів, що можуть бути змодельовані за величиною кореляції ознак. Прикладами таких алгоритмів кластеризації є CLIQUE та SUBCLU.

Загалом, на відміну від контрольованого машинного навчання, кластерний аналіз вказує на асоціації та закономірності в даних, але не говорить явно про те, що є ознакою закономірності та на що можуть вказувати ці закономірності даних.

До загальними етапів кластерного аналізу можна віднести:

- формулювання проблеми;
- підготовка та очищення набору даних;
- вибір ознак кластеризації;
- вибір функції відстані;
- вибір методу кластерного аналізу;
- визначення кількості кластерів;
- проведення аналізу;
- інтерпретація отриманих кластерів;
- оцінка точності кластеризації.

Ознаки, за якими слід проводити кластерний аналіз, слід вибирати, враховуючи проведені дослідження початкового набору даних, особливості предметної галузі, гіпотези, що перевіряються, та суб'єктивні судження дослідника. Також важливо обрати функцію відстані або подібності, що найкраще підходить до конкретної задачі (найчастіше використовується Евклідова відстань).

Застосування кластерного аналізу дуже широке, та охоплює велику кількість різноманітних областей знань, до яких можна віднести:

1) біологія та біоінформатика;
    1. екологія рослин і тварин — використовується для опису та для просторового та часового порівняння спільнот (сукупностей) організмів у неоднорідних середовищах;
    2. Genetic Генетична кластеризація людини — подібність генетичних даних використовується алгоритмами кластерного аналізу для визначення структур популяції;

2) медицина;
    1. медична візуалізація — розрізнення різних типів тканин у тривимірному просторі для різних застосувань;
    2. аналіз антимікробної активності — пошук закономірностей у стійкості мікробів до антибіотиків, класифікація антимікробних сполук ліків відповідно до їх механізму дії, класифікація антибіотики відповідно до їх антибактеріальної активності;
3) бізнес та маркетинг;
    1. дослідження ринку — розподіл загальної сукупності споживачів на сегменти ринку та краще розуміння взаємозв'язків між різними групами існуючих споживачів або потенційних споживачів;
    2. групування предметів покупок — групування усіх предметів покупок, доступних в Інтернеті, у набори унікальних продуктів;
4) всесвітня мережа;
    1. аналіз соціальних мереж — для виокремлення спільнот серед великих груп людей;
    2. групування результатів пошуку — у процесі інтелектуального групування файлів та веб-сайтів кластеризація може бути використана для створення більш коректного набору результатів пошуку порівняно із звичайними пошуковими системами, такими як Google;
5) суспільні науки;
    1. аналіз злочинності — кластерний аналіз може бути використаний для виявлення місцевостей, де частіше спостерігаються певні види злочинів;
    2. аналіз освітніх даних — використовується для виявлення груп шкіл чи учнів зі схожими ознаками (відсоток успішних учнів, ступіть комп'ютеризації тощо).

1.3 Огляд літератури

На теперішній час існує досить невелика кількість як електронної, так і паперової літератури щодо застосування кластерного аналізу у межах аналізу тональності тексту. Це пов’язано здебільшого з відносною новизною тематики, а також складністю семантичного аналізу тексту в цілому.

Застосування кластерного аналізу для з метою знаходження прихованих закономірностей у даних, в даному випадку семантичних забарвлень та думок, висловлених у тексті є підрозділом інтелектуального аналізу даних [1].

Загалом існує два основних академічних напрямки проведення аналізу тональності тексту: символічні методи (безконтрольні) та підходи за участю контрольованого машинного навчання (класифікації). Тож проаналізуємо існуючі роботи з цих підходів.

У дослідженні C. Cesaran та співавторів запроваджується найпростіший метод підрахунку балів слів, що входять до аналізуємого тексту [2], який полягає у опиті низки людей з метою надати оцінки документам, що виражають певну думку. Після надання оцінки людьми, для встановлення балів сформованих слів, у дослідженні застосовується стратегія оцінки слів з “псевдо-очікуваним” значенням, що походить від принципу очікуваних значень у математичній статистиці. Інша стратегія, яка називається оцінка прикметриків на основі “псевдо-стандартного-відхилення”, була представлена в тій же публікації. Вважається, що існує ще декілька інших стратегій отримання балів слів. Однак, оскільки всі ці функції ґрунтуються на людській інтуїції, вони недостатньо надійні через вплив на оцінку суб'єктивності, пов’язаної з індивідуальними особливостями освітнього та культурного досвіду різних людей. Крім того, системи, які значною мірою покладаються на людське втручання, є дорогими і низькомасштабованими, а тому непридатними для обробки великого обсягу даних.

У області символічних (безконтрольних) методів аналізу тексту є два напрямки досліджень. Перший полягає у класифікації великої кількості поглядів на біполярні орієнтації (позитивні чи негативні) [3]. Цей напрямок започаткували статті Pang B. зі співавторами [4] та Peter D. Turney зі співавторами [5]. Іншим напрямком дослідження є ідентифікація суб'єктивності/об'єктивності. Цей напрям зазвичай визначають як класифікацію досліджуваного тексту на один з двох класів: об’єктивний чи суб’єктивний [6].

WordNet – це лексична база даних для англійської мови. Вона групує англійські слова в набори синонімів та містить різні семантичні відношення між наборами синонімів. Розроблена Принстонським університетом у 1985 році база має на меті створити комбінований словник та тезаурус, що є більш інтуїтивно зрозумілими і які також підтримують аналіз тексту та алгоритми штучного інтелекту. Метод підрахунку слів за допомогою WordNet зосереджується лише на підрахунку прикметників. Оскільки WordNet визначає взаємозв'язок між синонімами, можна виміряти подібність або відстань між двома словами. Міра відстані основана на поняттях теорії графів. Усі прикметники в базі даних WordNet були зібрані та можуть бути об’єднані у граф. Вузлами графу є слова, а ребра виражають відношення синонімів. Відстань $d(w_i, w_j)$ між двома словами w_i і w_j – це довжина найкоротшого шляху між w_i і w_j. Якщо між ними немає шляху, відстань вважається нескінченною.

Для підрахунку прикметників в якості еталонних слів для вираження позитивних та негативних напрямків були обрані біполярні слова “добре” та “погано”. Тому для кожного прикметника w можна виміряти відстані $d(w, добре)$ та $d(w, погано)$. Вважається, що прикметники з меншою відстанню до “добре” є більш позитивними, а ті, що ближчі до “погано”, більш негативними.

Для формального визначення оцінки слова w існує наступний вираз:

$$EVA(w) = \frac{d(w, погано) - d(w, добре)}{d(добре, погано)} \quad (1.1)$$

Таким чином, слово *w* може отримати значення оцінки в інтервалі [-1, 1], де -1 виражає "погану" сторону, а 1 - "добру" сторону лексикону. Подібним чином метод можна також застосовувати для вимірювання прикметників за розмірами впливу та активності, встановлюючи такі референсні слова як “слабкий”, “сильний”, “позитивний” чи “активний”. Точність експерименту автора становить близько 70%. Це дослідження не застосовувало функції агрегування оцінок для підрахунку загальної оцінки текстів.

Стратегію оцінювання на основі веб-пошуку ввів Peter D. Turney [5]. Цей метод також вимагає вимірювання відстані між словами. Для визначення взаємозв'язку синонімічності слів у веб-документі Turney відкрив та розробив підхід для вимірювання подібності слів [7]. Підхід отримав назву “Pointwise Mutual Information – Information Retrieval” (PMI-IR). Він працює на припущенні, що терміни, які зустрічались разом, часто мають подібне значення. Оцінка “точкової взаємної інформації” між двома словами може бути виражена як:

$$PMI(w_1, w_2) = \log_2\left(\frac{p(w_1, w_2)}{p(w_1) * p(w_2)}\right) \qquad (1.2)$$

, де w_1 і w_2 – два слова; $p(w_i)$ – це ймовірність появи w_i (i = 1, 2), а $p(w_1, w_2)$ – ймовірність того, що w_1 і w_2 знаходяться поруч у тексті.

У практичних застосуваннях методу ймовірність появи термінів обчислюється шляхом підрахунку входжень (використовуючи систему AltaVista). Подібно до попередньо-описаного підходу, у поточному досліджені було обрано два референсні слова “відмінно” та “погано”.

Таким чином, що оцінку кожного терміна можна обчислити за формулою:

$$Оцінка(термін) = PMI(термін, \text{відмінний}) - PMI(термін, \text{поганий}) \qquad (1.3)$$

Доктор Turney проводив експерименти з розпізнавання тональності документів за допомогою бальних оцінок. Показник точності для чотирьох різних тем становив 84% (автомобілі), 80% (банки), 65,83% (фільми) та 70,53% (напрямки подорожей). Слід зазначити, що дані експерименту є незбалансованими . Близько 59% документів є позитивними, що означає, що завжди відгадування основного класу дасть точність ≥59%. Можна припустити, що рівень точності зменшиться при застосуванні цього методу до збалансованого набору даних.

Найвідоміші дослідження аналізу тональності з використанням контрольованого машинного навчання проводили Pang B. та співавтори. У 2002 році вони представили основні підходи до класифікації настроїв за даними оглядів фільмів [5] та представили метод вилучення об'єктивних речень що покращував результати попередніх експериментів у 2004 році [8]. Згодом, у 2005 році, вони розпочали роботу з присвоєння оцінки тональності документам за трибальною або чотирибальною шкалою [9]. Експериментальні дані складались із 700 оглядів фільмів із позитивною тональністю та 700 з негативною тональністю. Спочатку на цих документах була застосована методика обробки заперечень [10], щоб усунути похибку, спричинену запереченням слів. Для проведення класифікації даних було розглянуто три класифікатори: наївний Баєсовський класифікатор, класифікатор максимальної ентропії та метод опорних векторів (SVM). Ефективніст класифікаторів була провалідована в інших дослідженнях категоризації тексту і всі необхідні для дослідження попередньо-промарковані навчальні дані були зібрані. Для кожного алгоритму результати тренувань перевірялись шляхом перехресної перевірки. Для обробки експерименту були обрані різні ознаки класифікації, включаючи уніграми, біграми (два слова поспіль) та прикметники і т.п.. Рівень точності їх становив близько 80%, причому найвища точність досягала 82,9% (присутність уніграм із SVM), а найнижчий 72,8% (частота уніграм із SVM).

1.4 Системний аналіз методів кластерного аналізу

Аналіз ієрархій може слугувати інструментом системного аналізування алгоритмів кластеризації. Для цього перш за все необхідно визначити чіткі ознаки у відповідності до яких буде проводитися аналіз алгоритмів кластеризації:

- універсальність алгоритму;
- складність алгоритму;
- кількість необхідних даних;
- точність кластерного аналізу;
- необхідні об’єми обчислень.

Оскільки область комп’ютерних наук охоплює значну кількість алгоритмів кластеризації, дана кваліфікаційна робота буде оцінювати лише найпопулярніші з них з різних кластерних моделей:

- hierarchical clasterization (модель зв’язаності);
- k-means (центроїдна модель);
- c-means (нечітка кластеризація).

Також визначимо три рівня аналізу ієрархій алгоритмів кластеризації:

- множини альтернатив;
- критерії порівняння методів;
- вибір методу розв'язання.

Спираючись на отримані критерії визначення кращого методу кластерного аналізу сформуємо матрицю парних порівнянь [11]. Попарне порівняння можна охарактеризувати як порівняння об’єктів чи сутностей за певними ознаками чи якостями по парах з метою визначення переваги однієї сутності над іншою, або ж їхньої тотожності. Даний метод порівняння набув широкого розповсюдження у економічному аналізі, системах штучного інтелекту, психології, теорії прийняття рішень.

Матриця парних порівнянь надає можливість математичного порівняння ознак алгоритмів кластерного аналізу. Важливість (або вагу) тієї чи іншої ознаки можна визначити на основі переваги цієї ознаки над іншою. На основі важливості ознак далі можна визначити відношення відповідності і індекси критеріїв, що дають змогу оцінити ступінь узгодженості обчислень, а також знайти вектори локальних пріоритетів.

Значення узгодженості <= 0.2 вказує на коректність і правильність підібраних критеріїв порівняння, в той час як більші значення кажуть про необхідність перегляду ознак порівняння та підбору більш коректних.

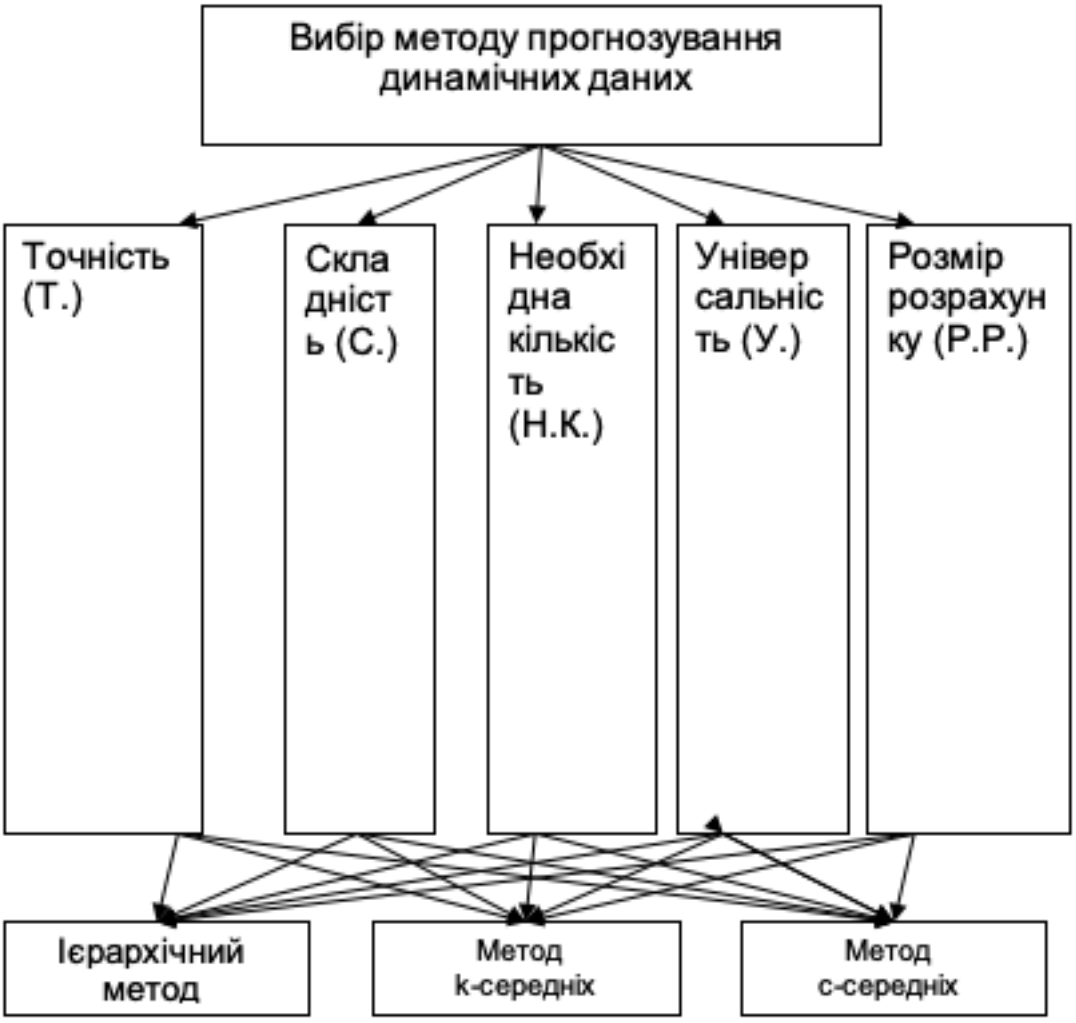

Рисунок 1.1 – Схема ієрархії критеріїв порівняння методів кластерного аналізу

У відповідності до ієрархічної структури критеріїв на рисунку 1.1 можна скласти наступну матрицю парних порівнянь:

Таблиця 1.1 – Матриця парних порівнянь 1-го рівня

Кри-терій	Точ н.	Склад.	Необх. к-ть	Універс альн.	Розм. Розр.	Власний вектор	Вектор пріоритетів
Точн.	1	$\frac{1}{5}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	2,605	0,427
Склад.	5	1	2	3	$\frac{1}{2}$	0,582	0,095
Необх. к-ть	2	$\frac{1}{2}$	1	2	$\frac{1}{2}$	1	0,165
Універ сальн.	3	$\frac{1}{3}$	$\frac{1}{2}$	1	$\frac{1}{3}$	1,431	0,234
Розм. Розр.	4	2	2	3	1	0,461	0,079

З метою розрахунку індексу узгодженості необхідно спочатку знайти власний вектор кожного критерію за формулою:

$$BB=\left(\prod_{i=1}^{k} K_{ij}\right)^{\frac{1}{k}} \tag{1.4}$$

, де i – індекс рядка, j – індекс. стовбця, k – кількість рядків, K – судження

Також необхідно знайти максимальне власне значення кожного критерію за формулою:

$$\lambda_{max}=\sum_{i=1}^{k}\left(\sum_{j=1}^{m} K_{ij} * ВП\right) \tag{1.5}$$

, де i – індекс рядка матриці, j – індекс стовбця, k – кількість рядків, m – кількість стовбців, K – судження, ВП — значення компоненти вектора

Результуючий індекс узгодженості розраховується за формулою:

$$ІУ = \frac{\lambda_{max} - n}{n - 1} \quad (1.6)$$

, де λ_{max} – максимальне особисте значення, n – кількість критеріїв

Результуюче відношення узгодженості розраховується за формулою:

$$ВУ = \frac{ІУ}{ВипУ} \quad (1.7)$$

, де $ВипУ$ – випадкова узгодженість, n – кількість критеріїв матриці

Оцінка індексу узгодженості:

$$ІУ = \frac{5,191 - 5}{5 - 1} = 0.048$$

Оцінка відносної узгодженості:

$$ВУ = \frac{0,048}{1,12} = 0,043$$

Виходячи з отриманих індексу узгодженості та відносної узгодженості можна стверджувати про коректність заповнення матриці та про те, що твердження не суперечать один одному. Вектор локальних пріоритетів має наступний вигляд:

$$\vec{p}^{K} = (0,427;\ 0,095;\ 0,165;\ 0,234;\ 0,079)$$

Далі з метою оцінки альтернатив необхідно провести якісний аналіз методів кластерного аналізу за обраними критеріями оцінювання. Таблиці 1.2 –

1.6 є матрицями парних порівнянь, що дають змогу зробити висновки про доречність вибору того чи іншого алгоритму кластеризації.

Таблиця 1.2 – Матриця парних порівнянь критерію «Точність»

Точність	“hierarch. clustering”	“k-means”	“c-means”	Власний вектор	Вектор пріоритетів
“hierarch. clustering”	1	$\frac{1}{5}$	$\frac{1}{2}$	2,154	0,598
“k-means”	5	1	2	0,464	0,127
“c-means”	2	$\frac{1}{2}$	1	1	0,275

Оцінка індексу узгодженості:

$$ІУ = \frac{3{,}0015 - 3}{3 - 1} = 0.00075.$$

Оцінка відносної узгодженості:

$$ВУ = \frac{0{,}00075}{0{,}58} = 0.0013.$$

Результуючий вектор локальних пріоритетів:

$$\vec{p}_1^{\,A} = (0{,}598;\ 0{,}127;\ 0{,}275)$$

Таблиця 1.3 – Матриця парних порівнянь критерію «Складність»

Складн.	“hierarch. clustering”	“k-means”	“c-means”	Власний вектор	Вектор пріоритетів
“hierarch. clustering”	1	5	2	0,464	0,127
“k-means”	$\frac{1}{5}$	1	$\frac{1}{2}$	2,154	0,598
“c-means”	$\frac{1}{2}$	2	1	1	0,275

Оцінка індексу узгодженості:

$$ІУ = \frac{3{,}0015 - 3}{3 - 1} = 0.00075$$

Оцінка відносної узгодженості:

$$ВУ = \frac{0{,}00075}{0{,}58} = 0.0013$$

Результуючий вектор локальних пріоритетів:

$$\vec{p}_2^{\,A} = (0{,}127;\ 0{,}598;\ 0{,}275)$$

Таблиця 1.4 – Матриця парних порівнянь критерію «Кількість даних»

Кільк. дан.	“hierarch. clustering”	“k-means”	“c-means”	Власний вектор	Вектор пріоритетів
“hierarch. clustering”	1	2	2	0,63	0,22
“k-means”	$\frac{1}{2}$	1	1	1,26	0,39
“c-means”	$\frac{1}{2}$	1	1	1,26	0,39

Оцінка індексу узгодженості:

$$ІУ = \frac{3 - 3}{3 - 1} = 0$$

Оцінка відносної узгодженості:

$$ВУ = \frac{0}{0{,}58} = 0$$

Результуючий вектор локальних пріоритетів:

$$\vec{p}_3^{\,A} = (0{,}22;\ 0{,}39;\ 0{,}39)$$

Таблиця 1.5 – Матриця парних порівнянь критерію «Універсальність»

Універс.	“hierarch. clustering”	“k-means”	“c-means”	Власний вектор	Вектор пріоритетів
“hierarch. clustering”	1	$\frac{1}{7}$	$\frac{1}{4}$	3,036	0,705
“k-means”	7	1	3	0,362	0,083
“c-means”	4	$\frac{1}{3}$	1	0,91	0,212

Оцінка індексу узгодженості:

$$ІУ = \frac{3{,}029 - 3}{3 - 1} = 0{,}015$$

Оцінка відносної узгодженості:

$$ВУ = \frac{0{,}015}{0{,}58} = 0{,}025$$

Результуючий вектор локальних пріоритетів:

$$\vec{p}_4^{\,A} = (0{,}705;\ 0{,}083;\ 0{,}212)$$

Таблиця 1.6 – Матриця парних порівнянь критерію «Об’єм розрахунків»

Об. розрах.	“hierarch. clustering”	“k-means”	“c-means”	Власний вектор	Вектор пріоритетів
“hierarch. clustering”	1	$\frac{1}{2}$	$\frac{1}{3}$	1,817	0,534
“k-means”	2	1	4	0,5	0,145
“c-means”	3	$\frac{1}{4}$	1	1,1	0,321

Оцінка індексу узгодженості:

$$ІУ = \frac{3{,}042 - 3}{3 - 1} = 0{,}021$$

Оцінка відносної узгодженості:

$$ВУ = \frac{0{,}021}{0{,}58} = 0{,}036$$

Результуючий вектор локальних пріоритетів:

$$\vec{p}_5^{\,A} = (0{,}534;\ 0{,}145;\ 0{,}321)$$

Далі, виходячи з отриманих значень узгодженості та локальних пріоритетів обраних критеріїв, можна побудувати вектор глобальних пріоритетів. Слід зазначити, що індекс узгодженості дорівнює сумі скалярного добутку векторів локальних пріоритетів з вектором індексів узгодженості критеріїв та індексу першого рівня.

Таблиця 1.7 – Матриця остаточних результатів задачі вибору методу кластерного аналізу

Критерій / Альтернатива	Розм. Розр.	Універс.	Необх. кіл-ть	Складн.	Точн.	Вектор пріоритетів
“c-means”	0,321	0,212	0,39	0,275	0,275	0,283
“k-means”	0,145	0,083	0,39	0,598	0,127	0,2
“hierarch. clustering”	0,534	0,705	0,22	0,127	0,598	0,76

Перевіривши отримані дані з матриці 1.7 можна зробити висновок про те, що вони задовільняють критерії узгодженості. Згідно з даними, оптимальним методом кластерного аналізу є “Hierarchical clustering”, оскільки йому належить максимальне значення вектору пріоритетів.

Після обрання методу кластерного аналізу згідно методики аналізу ієрархій необхідно проаналізувати існуючі проблеми обраних методів. Для цього треба визначити небажані, бажані та критичні критерії кожного методу. Виходячи з теоретичної інформації щодо алгоритмів кластеризації одним з найбільших небажаних критеріїв є висока потреба у програмних ресурсах.

Аби розв’язати вищезазначену проблему необхідно провести кластеризацію критеріїв системи, у кожного я яких є якості розглянутого методу моделювання. Нижче представлені категорії вдоволеності на основі критерії методів кластерного аналізу з описом характеристик кожної з них:

- небажані якості — занадто високий рівень затрат на розрахунки та неточності чи похибки обчислень;
- бажані якості — висунення припущень щодо функціонального та психічного станів з високою точністю та помірною похибкою;
- критичні якості — здебільшого це універсальність та складність програмної реалізації методу кластерного аналізу.

Далі побудуємо структуру ієрархій спираючись на описані категорії вдоволеності критеріїв методів кластерного аналізу:

– цілі, у вигляді проблеми незадоволеності;
– класифікація незадоволення;
– характеристики незадоволень.

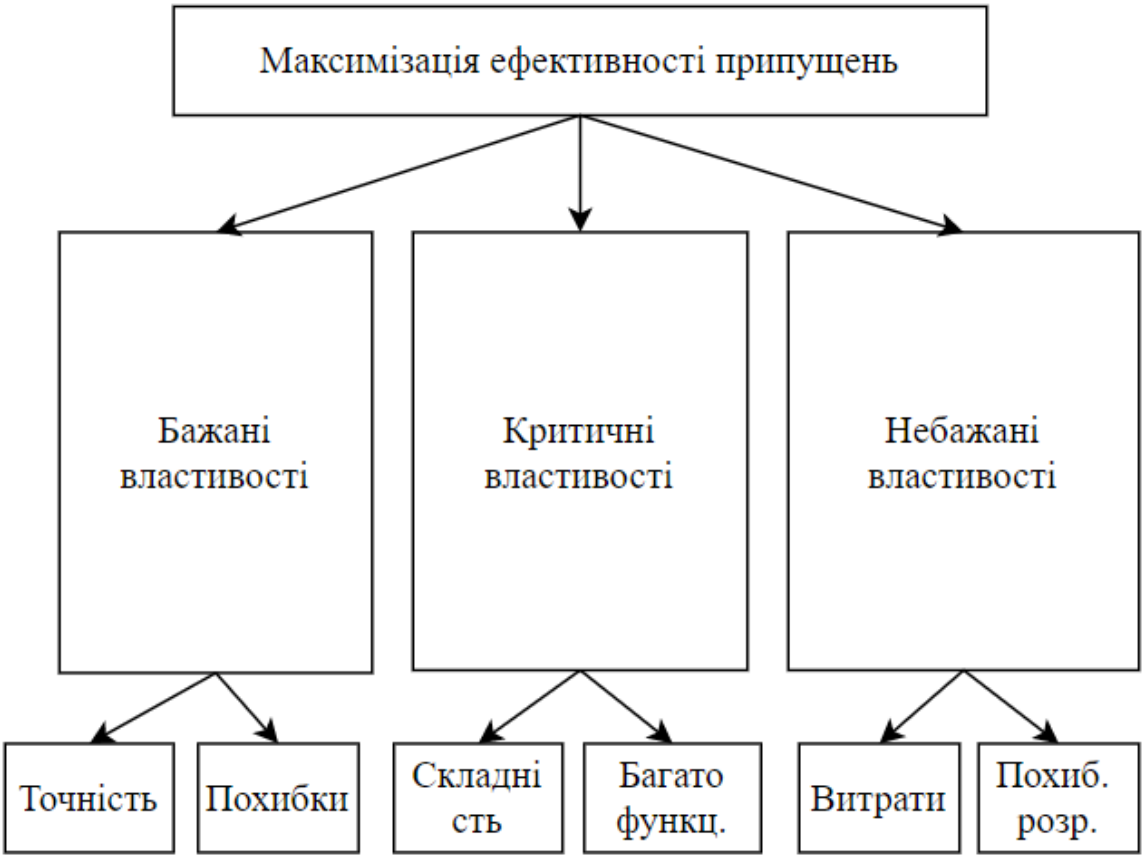

Рисунок 1.2 – Схема структури ієрархії категорій вдоволеності

Спираючись на розроблену структуру ієрархії максимізації ефективності висунення припущень [12], можна зробити висновок, що модель є закінченою.

1.5 Вектор пріоритетів вдоволеності методом ієрархічного аналізу

Матриця парних порівнянь першого рівня надасть можливість визначити оцінку глобального вектора пріоритетів. Таку ж матрицю потрібно побудувати для кожної з категорій вдоволеності.

Таблиця 1.8 – Матриця парних порівнянь 1-го рівня

Категорія	Бажані власт.	Критичні власт.	Небажані власт.	Власний вектор	Вектор пріоритетів
Бажані власт.	1	$\frac{1}{4}$	$\frac{1}{2}$	2	0,557
Критичні власт.	4	1	3	0,437	0,123
Небажані власт.	2	$\frac{1}{3}$	1	1,145	0,319

Далі необхідно побудувати матриці парних порівнянь для небажаних якостей, для критичних якостей та бажаних якостей, згідно до матриці першого рівня.

Таблиця 1.9 – Матриця парних порівнянь категорії “Бажані якості”

Бажані якості	Точність	Пох. прогноз.	Власний вектор	Вектор пріоритетів
Точність	1	$\frac{1}{3}$	1,732	0,751
Пох. прогноз.	3	1	0,577	0,246

Таблиця 1.10 – Матриця парних порівнянькатегорії “Критичні якості”

Критичні якості	Складн.	Багатофункц.	Власний вектор	Вектор пріоритетів
Складн.	1	5	0,447	0,165
Багатофункц.	$\frac{1}{5}$	1	2,236	0,834

Таблиця 1.11 – Матриця парних порівнянь категорії “Небажані якості”

Небажані якості	Витрати	Похиб. результатів	Власний вектор	Вектор пріоритетів
Витрати	1	3	1,732	0,751
Похиб. результатів	$\frac{1}{3}$	1	0,577	0,248

Після побудування матриць парних порівнянь та переконання у вірності значень треба сформувати вектор глобальних пріоритетів (рисунок 1.3) у вигляді діаграми.

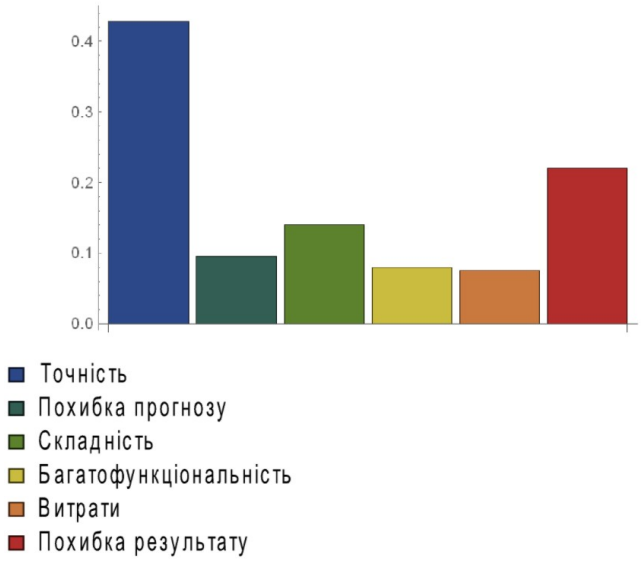

Рисунок 1.3 – Схема глобальних пріоритетів

Відповідно до отриманих результатів є можливість стверджувати про особливу важливість таких ознак, як зниження похибки прогнозу, зниження складності програмної реалізації, а також підвищення точності результатів. Таким чином більшість зусиль покладатимуться на розв’язання даних задач.

1.6 Загальна постановка задачі

Для коректного проведення кластерного аналізу, а саме підбору параметрів моделі, велике значення має процес обрання алгоритму аналізу оброблюваних даних в залежності від характеристик цих даних. Використовуючи методи порівняння моделі можна обрати найбільш точний алгоритм аналізу вхідних даних.

Загалом дані користувачів соціальних мереж зберігаються у неструктурованому виді, здебільшого як набори символів у певних кодуваннях. Всі ці дані потенційно здатні відображати функціональний та психічний стан того чи іншого користувача. Через неоднорідність даних виникає проблема кластерного аналізу таких даних програмою, та необхідність їх перетворення у структуроване, здебільшого числове представлення, за допомогою алгоритмів попередньої обробки даних. Можна сформувати такі етапи розв’язання даної проблеми:

- розробити програму для отримання даних користувачів та їх перетворення у нормалізоване представлення;
- трансформувати текстові дані у загальне нормалізоване представлення (цифрове);
- розробити програму для проведення кластерного аналізу;
- розробити програму для відображення результатів;
- провести кластерний аналіз даних.

1.7 Постановка задач дослідження

Згідно з результатами системного аналізу визначимо основні задачі дослідження у кваліфікаційній роботі:

- побудова моделі парних порівнянь у відповідності до обраних критеріїв методу кластерного аналізу;
- проведення аналізу існуючих методів кластерного аналізу;
- як результат аналізу у відповідності до моделі обрати найбільш ефективний метод кластерного аналізу;
- здійснити кластерний аналіз даних за допомогою реалізованої програми;
- провести аналіз результатів;
- зробити висновки.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

2.1 Метод "k-means"

Даний метод відноситься до групи "ітераційних" методів кластерного аналізу. Метод вимагає заздалегідь визначити число кластерів на котрі будуть розподілені дані. Одна ітерація алгоритму складається з двох кроків: знаходження центру кластеру (центроїд) для наявних у кластері даних, а потім розподілення вакантних даних до кластерів у відповідності до щойно розрахованих центроїдів. Вибір кластеру для певної одиниці даних відбувається в залежності від відстані між одиницею та центроїдом. Найпопулярніша формула для знаходження відстані — це Евклідова відстань:

$$p(x,y)=\|x-y\|=\sqrt{\sum_{p=1}^{n}\left(x_p-y_p\right)^2} \qquad (2.1),$$

де $x,y\in R^n$ – місцезнаходження даних на координатній площині. Треба взяти до уваги перелік спостережень $\left(x^{(1)},x^{(2)},\ldots,x^m\right), x^{(m)}\in R^n$. Алгоритм k-means розподіляє k спостережень (кластерних точок) на m кластерів. Здебільшого виконується властивість $m\le k, S=\left\{S_1,S_2,\ldots,S_m\right\}$. Головна задача даного методу — мінімізація суми квадратичного відхилення точок кластерів від центрів цих кластерів — описується наступною формулою:

$$min\left[\sum_{i=1}^{k}\sum_{x\in S_i}\left\|x^{(j)}-\mu_i\right\|^2\right] \qquad (2.2),$$

де μ_i – центроїд кластера S_i. Цей метод мінімізації називається "методом головних точок", де головні точки — це центри кластерів, і дозволяє досягти найкращої апроксимації даних.

Першим кроком методу k-means є визначення числа кластерів розробником, котрий мусить покладатися на власні знання специфіки даних, закономірностей їх розташування, ключових характеристик та відмінностей. Метод k-means може показувати досить низьку ефективність кластеризації у разі обрання таких якостей даних, що не характеризують їх вичерпно. В такому разі дані будуть групуватись занадто близько.

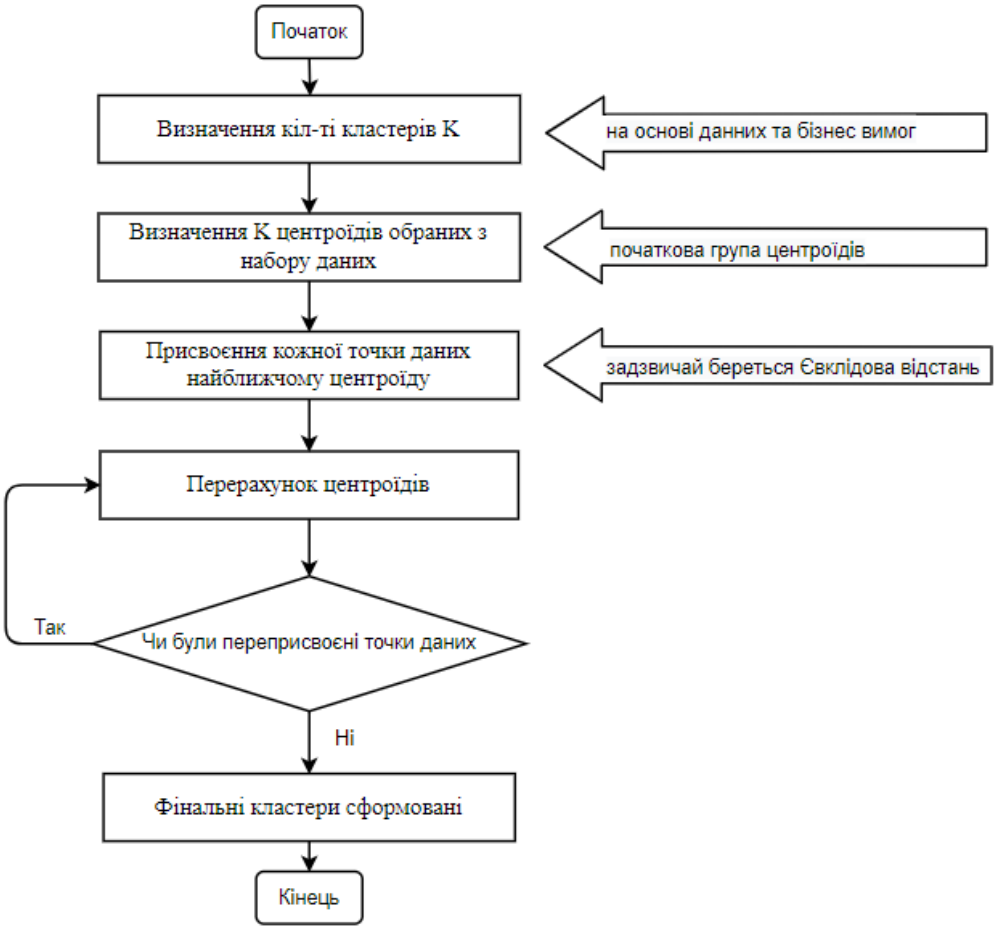

Рисунок 2.1 — Алгоритм “k-means”

Наступним етапом після знаходження числа кластерів є вибір центроїдів кластерів — випадкових точок на площині. Проте існує оптимізація цього етапу, що дозволяє досягти більшої точності та швидкості кластеризації, вона полягає у виборі точок-центрів в залежності від найбільшої можливої

початкової відстані між вірогідними центрами кластерів. Такий алгоритм має назву “k-means++” та зменшує число ітерацій.

Розрахунок центроїду кожного з кластерів відбувається за наступною формулою:

$$\mu_i = \frac{1}{S_{i\, x^{(j)} \in S_i}} \sum x^{(j)} \tag{2.3}$$

Таким чином ітеративний розрахунок центроїдів кожного з кластерів з поетапним збільшенням точності розташування центрів кластерів — це основне завдання методу “k-means”. Розробник визначає як число кластерів так і число ітерацій алгоритму.

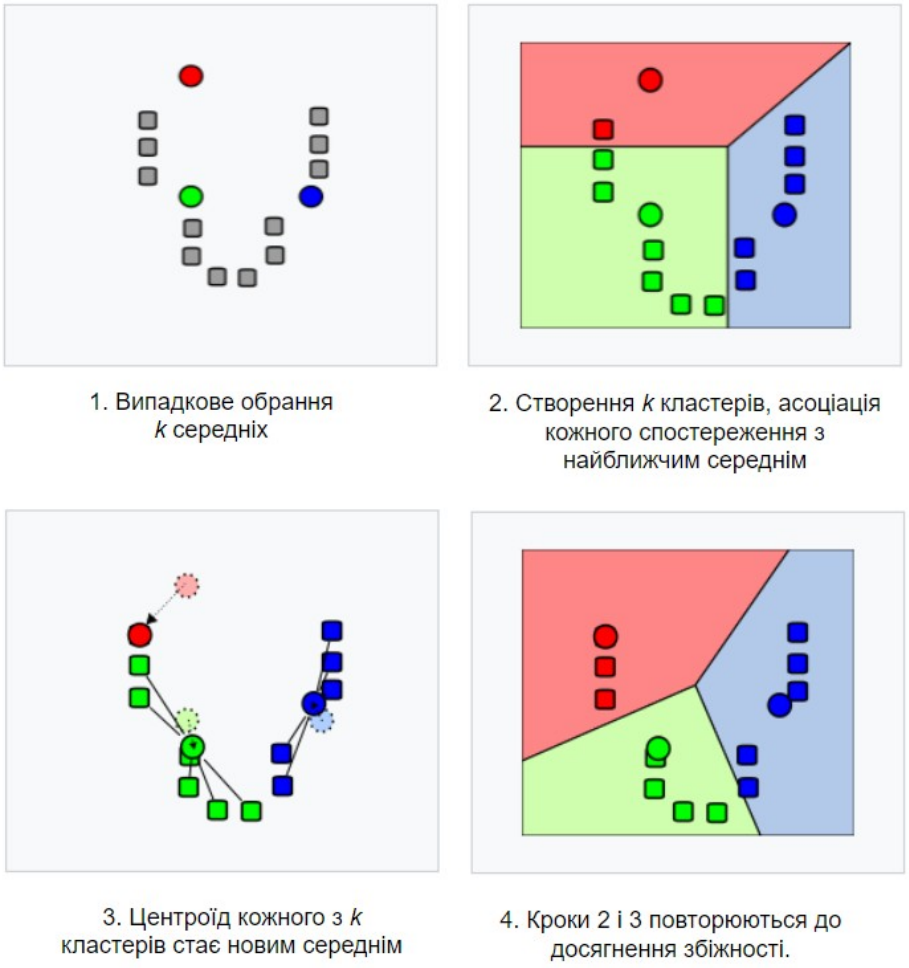

Рисунок 2.2 — Демонстрація алгоритму “k-means”

Є також модифікація алгоритму, коли кількість ітерацій не задається, в такому випадку зупинка відбувається програмно коли координати центрів не змінюються, або розробник зупиняє алгоритм вручну.

Проте бувають випадки, при яких дані настільки розрізнені, що до початку роботи алгоритму відсутнє навіть незначне формування у кластери. У такому разі досить складно точно визначити необхідне число кластерів за видимими закономірностями розташування даних. Такі випадки вказують на чи не найголовніший недолік методу “k-means” – необхідність ручного задання числа кластерів. Для даних такого типу потрібно застосовувати методи кластерного аналізу, що не мають таких недоліків.

До інших недоліків алгоритму можна віднести:

– точність кластеризації залежить від вибору початкових центроїдів, проте їх оптимальний вибір невідомий;
– в той час як алгоритм гарантує досягнення одного з локальних мінімумів суми квадрата відхилення, проте не завжди вдається досягти глобального мінімуму.

2.2 Метод “c-means”

Основним недоліком “k-means”, про який вже було згадано, є потреба у заданні числа ітерацій та кластерів розробником, що інколи призводить до неочевидної та небажаної поведінки алгоритму.

Так, при перевищенні кількості визначених розробником ітерацій, алгоритм закінчує роботу, що не гарантує досягнення необхідної точності кластеризації.

Рисунок 2.3 — Алгоритм “c-means”

Також можлива ситуація особливого розташуванням даних — на перетині кластерів. В такому випадку метод “k-means” може працювати нескінченно довго, адже алгоритм не може остаточно визначити приналежність

одиниці даних то того чи іншого кластера і переносить дані між двома кластерами у обидва напрямки.

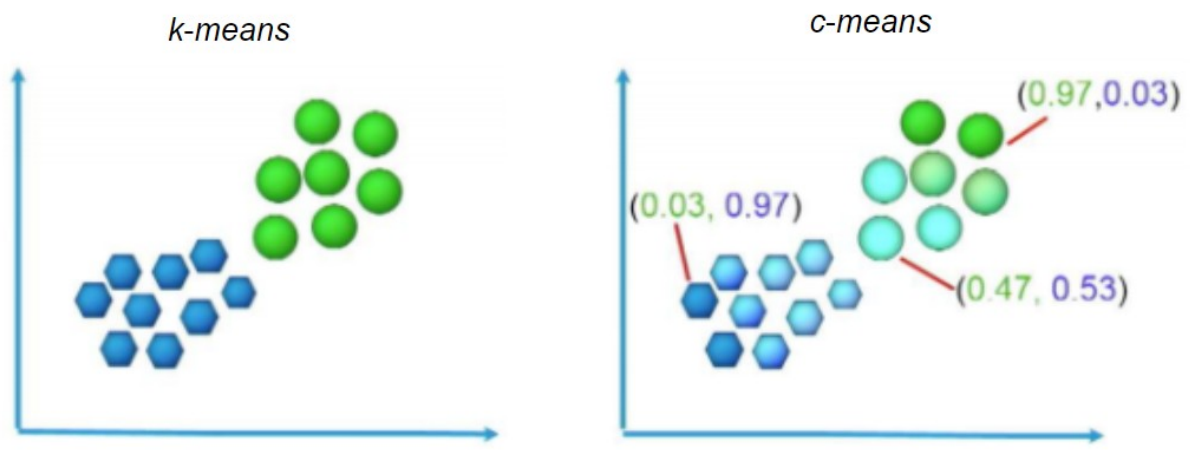

Рисунок 2.4 — Порівняння методів “k-means” та “c-means”

Алгоритм кластерного аналізу “c-means” є імовірнісним, що дозволяє розв’язати проблему належності даних шляхом знаходження імовірності приналежності даних до кожного кластеру (ймовірнісна або нечітка кластеризація). Задача мінімізації відстані до центроїду вирішується за допомогою функції:

$$E = \sum \sum u_{ij}^{m} \cdot \left\| x_i - c_j \right\|^2 \qquad (2.4),$$

де $\sum_j u_{ij} = 1, j = 1 \ldots p$.

2.3 Метод “Hierarchical clustering”

Взагалі алгоритми кластеризації можна поділити на дві групи. До першої належать так звані “зовнішні”, дані для таких методів повинні мати лише одну

ключову ознаку, за якою проводиться кластеризація. До другої групи належать “внутрішні” алгоритми, що не накладають обмежень на дані і дозволяють мати декілька рівнозначних критеріїв, згідно з якими проводиться кластерний аналіз.

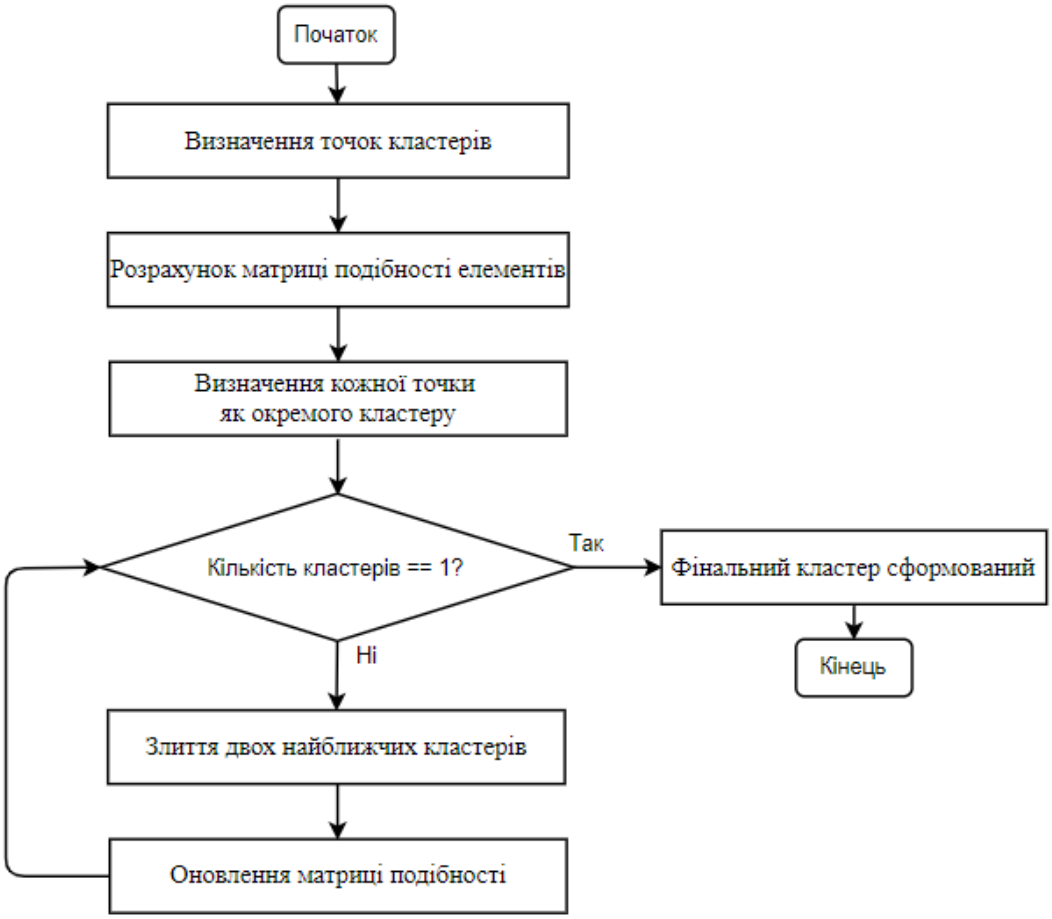

Рисунок 2.5 — Алгоритм ієрархічної кластеризації

До “внутрішніх” алгоритмів можна віднести як ієрархічні, так і неієрархічні [12], перша підгрупа методів має у основі древовидну структуру, в той час як друга задзвичай позбавлена чіткої структури, або структура досить комплексна. Для ієрархічних методів кожна одиниця даних є окремим класом. Відстань від одиниці даних до центроїду кластери розраховується за допомогою наступної функції:

$$R(\{x\},\{x'\}) = p(x,x'), \tag{2.5}$$

На кожній ітерації методу знаходяться найбільш близькі за відстанню кластери L та K, з яких формується спільний кластер $M = L \cup K$.

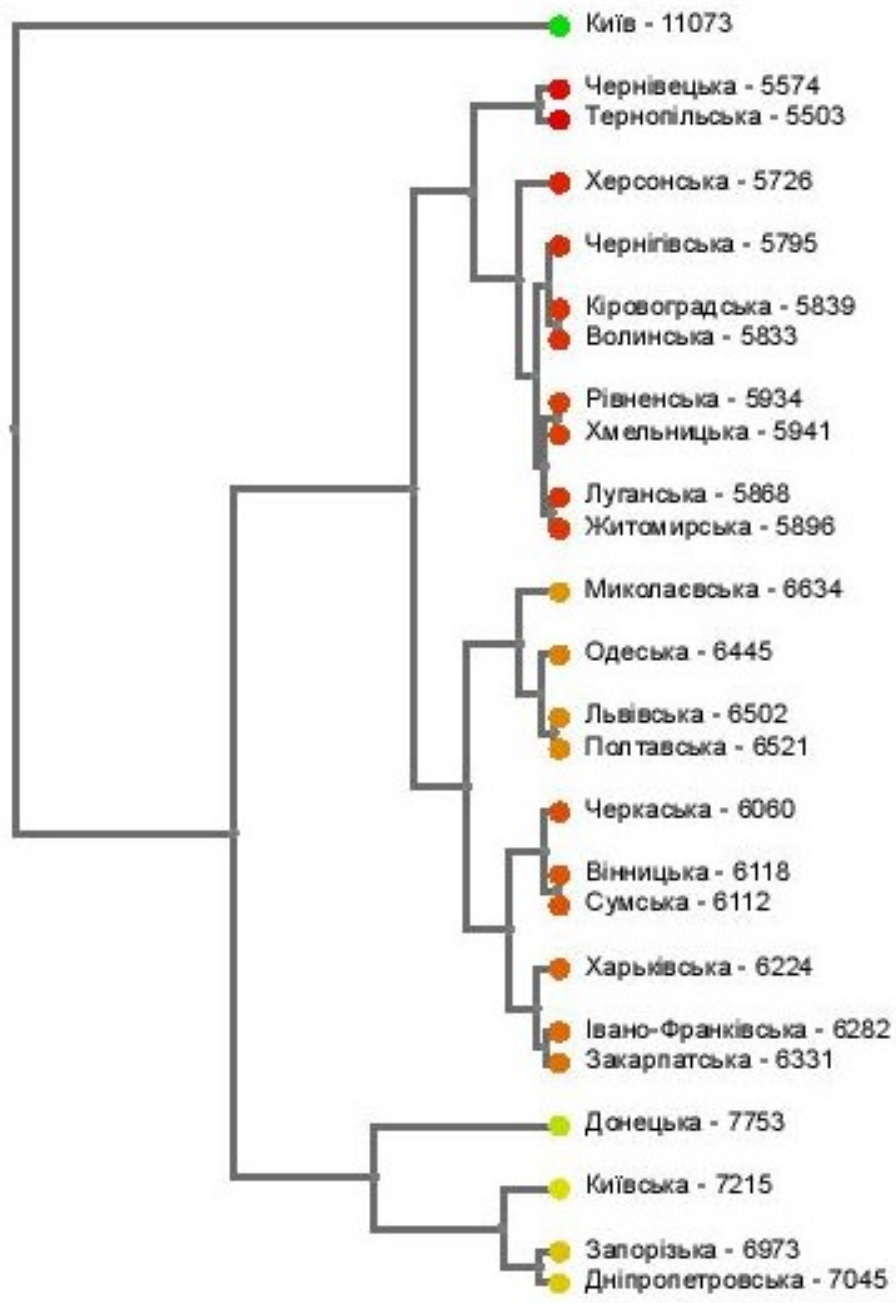

Рисунок 2.6 — Приклад ієрархічної кластеризації областей України за середньою заробітньою платнею

Всі ітерації алгоритму з пари найближчих один до одного кластерів L та K утворює інший кластер $M = L \cup K$. За допомогою відстаней $R(L,K), R(L,N), R(K,N)$, що знаходяться до формування кластеру $M = L \cup K$, можна знайти відстань від сформованого кластеру M до довільного кластера N.

Зазначені відстані до кластера N розраховуються за допомогою наступної формули:

$$R(L \cup K, N) = \alpha_{\cup} R(L,N) + \alpha_{\cup} R(K,N) + \beta R(L,K) + \gamma R(L,N) - R(K,N), \qquad (2.6)$$

, де $\alpha_{\cup}, \beta, \gamma$ – це цифрові параметри. У даній формулі відстань від одного кластера до іншого (вузли дерева) є дугою дерева.

Не дивлячись на недолік даного алгоритму кластерного аналізу, що вимагає вибір різних за сутністю ознак, не схожих одна на одну, даний метод показує високу ефективність та точність кластеризації даних за рахунок можливості перегляду сформованих кластерів на кожному кроці. Тобто, алгоритм передбачає повернення на попередній крок, до злиття кластерів, у тому разі, коли відстань між сформованими кластерами занадто велика [13].

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Словник тональності

Ключовим аспектом аналізу тональності, також відомого як “sentiment analysis” або семантичний аналіз або аналіз настроїв, є аналіз тексту для розуміння думки, яку автор намагався висловити цим текстом [14]. Як правило, ми оцінюємо ці настрої позитивним чи негативним значенням, що називається полярністю. Загальні настрої часто приймаються як позитивні, нейтральні або негативні за знаком оцінки полярності. Існує два основних підходи до аналізу настроїв:

- контрольоване машинне навчання або методи глибокого навчання;
- безконтрольні підходи, засновані на лексиконі;

Для першого підходу нам зазвичай потрібні попередньо позначені, або “промарковані” дані. Другий підхід покладається на словник слів, попередньо оцінених спеціальним алгоритмом машинного навчання. Саме другий підхід і буде використаний в даній кваліфікаційній роботі.

В рамках даної кваліфікаційної роботи сентимент-аналіз дозволить визначити загальну емоційну навантаженість постів користувача у соціальній мережі. З метою проведення аналізу настроїв, що в свою чергу дає можливість оцінити психічний та функціональний стан користувача, буде використовуватися словник AFINN, назва якого розшифровується як “Affective Norms for English Words”, тобто “Емоційні норми слів англійського алфавіту”.

Словник адаптований під усі розповсюджені мови програмування у вигляді зовнішніх програмних бібліотек, що надають програмний інтерфейс роботи зі словником, тим самим приховують зайву логіку та спрощують складність взаємодії. До прикладу бібліотека автоматично парсить дані користувача та форматує їх до стандарту, що підтримується AFINN.

Данні цього словника тональності зберігаються у файлі операційної системи у вигляді пар ключ-значення, де ключ — це слово, а значення — це оцінка емоційної забарвленості слова. Оцінка тональності має діапазон (-5; 5) балів, де негативно-забарвлені слова розташовуються на від'ємній частині шкали, нейтральні мають оцінку нуль, а позитивно-забарвлені на додатній частині шкали.

3.2 Дані для аналізу тональності постів

Для об'єктивності результатів кластерного аналізу потрібна досить велика вибірка даних користувачів соціальних мереж, тобто потрібен такий об'єм даних, при якому проведений кластерний аналіз покаже достатньо високу точність.

Twitter – це мікроблогінг сервіс де користувачі можуть відправляти та читати текстові пости, що називаються "твіти". Твіти це короткі повідомлення, що в довжину можуть досягати 140 символів. Користувачі мають змогу підписуватись на інших користувачів аби отримувати їх твіти як тільки ті будуть створені. На ці твіти потім можна дати відповідь, розпочинаючи таким чином діалог, або "ретвітнути", в такому випадку користувач постить оригінальний твіт згадуючи автора твіта з метою поділитися запозиченим твітом зі своїми підписниками. Twitter дуже популярний у всьому світі і на даний момент містить більше ніж півміл'ярда користувачів.

Twitter дозволяє шукати чужі твіти. Користувач може додати символ "#" у рядку пошуку аби знайти твіти за ключовим словом або конкретною тематикою. Слова з таким символом називаються "хеш-теги" та слугують посиланнями на колекцію твітів з подібною тематикою. Програмне

забезпечення. Twitter слідкує за популярними словами та хеш-тегами, таким чином спрощуючи користувачам пошук трендових новин.

У кінці 2010 року світом промайнула серія демонстрацій та протестів проти диктатури режимів що почалася з країн Ближнього Сходу та мала назву “Арабська Весна”. Під час цих подів Twitter став популярним засобом для комунікації серед протестувальників, що використовували соціальну мережу з метою організації протестів та як засіб висловлювання власної думки котру могли почути люди по всьому світу. Також Twitter часто використовується для організації допомоги та поширення інформації під час природних катастроф, таких як землетруси чи паводки.

Підсумовуючи вищесказане, мережа Twitter є надзвичайно цінним джерелом даних для проведення аналізу тональності в рамках даної кваліфікаційної роботи, адже кожної секунди близько дев’яти тисяч твітів з усіх можливих куточків світу відправляється користувачами мережі. Цей потік даних містить думки різних як за соціальним статусом так і за життєвими поглядами чи інтересами людей на будь-які теми, що структуровані завдяки зручному для аналізу формату збереження даних користувача у вигляді текстових постів. Також мережа надає потужне програмне API для роботи з цими даними [15].

Як і з багатьма іншими програмними API для його використання необхідно спочатку зареєструватися як розробник, а потім зареєструвати програмний додаток, для того аби саме цей додаток мав унікальний доступ до API мережі. З метою отримання ключів автентифікації додатку необхідно пройти додаткову верифікацію. Подібна схема є і у Google Cloud, де для доступу до програмних API сервісів необхідно створювати аккаунт сервісу, або “service account”.

В залежності від підключеної підписки користувачу API буде доступний різний перелік функцій.

Тип підписки “Standard” є безоплатним та надає доступ до базових функцій роботи з API, таких як:

- вибірка постів користувача за останній календарний тиждень;
- програмне створення та видалення як твітів, так і приватних повідомлень;
- фільтрування твітів користувача за певними критеріями (дата, кількість уподобань, кількість розповсюджень і т.п.);

Також варта уваги “Enterprise” підписка, що є платною, проте надає набагато ширший перелік функцій роботи з програмним API, таких як:

- збільшений ліміт давності твітів — до одного календарного місяця;
- розширений доступ мета-інформації твітів, наприклад до хештегів, посилань розповсюджень, згадування, гео-інформація;
- збільшений ліміт кількості вибірки твітів в межах одного запиту до API;
- фільтрування твітів користувача за певними критеріями (дата, кількість уподобань, кількість розповсюджень і т.п.);
- збільшена швидкість відгуку запитів до програмного API, збільшений ліміт тротлінгу;
- можливість виконання пошукових запитів у SQL-like форматі;
- можливість оптимізації вичитки даних за рахунок використання batch-запитів;

Подібно до Deep Archive у сервісі Amazon S3, через значні об’єми даних що оброблюються мережею щодня, Twitter оптимізує збереженя даних користувача шляхом архівації даних старших одного календарного місяця, таким чином досягається зменшення обсягу фізичної пям’яті, зайнятої даними. Процес запиту таких даних дещо складніше від вибірки свіжих даних, та вимагає залучення більшої кількості апаратних ресурсів для відтворення даних з архіву.

Від типу запитуваних даних залежить вартість запиту до програмного API мережі Twitter.

Для задач даної кваліфікаційної роботи функціоналу API “Standard” підписки буде достатньо. З метою взаємодії з програмним API мережі Twitter та відображення результатів роботи у зручному графічному інтерфейсі користувача будуть використані засоби мови програмування C#, а саме бібліотека Microsoft.Owin.Security.Twitter для автентифікації та бібліотеки HttpClient і Json.NET для роботи безпосередньо з API.

3.3 Розробка програми для аналізу тональності постів

З метою розробки програми для аналізу тональності постів користувачів мережі Twitter перш за все необхідно було визначити мову програмування, основними критеріями вибору стали: наявність можливості реалізації графічного інтерфейсу користувача, відносна легкість освоєння, популярність мови та наявність вичерпної документації.

Таким чином була обрана мова C#, що є об’єктно-орієнтованою, має строгу статичну типізацію, а також синтаксис схожий на Java та C++, адже є C-подібною мовою, підтримує поліморфізм, властивості, коментарії, події та інші важливі структури мов програмування. Мова має високу популярність, велике ком’юніті користувачів та безліч якісних допоміжних програмних бібліотек.

Також мова програмування C# має декілька бібліотек для створення графічного інтерфейсу користувача, а саме десктоп-програм, таких як Windows Forms, WPF та UWP. Для даної кваліфікаційної роботи була обрана бібліотека Windows Forms, основним чинником вибору стала її простота використання, наявність широкого вибору елементів інтерфейсу та велика кількість документації і туторіалів.

3.4 Розробка програми для кластеризації та її візуалізації

З метою розробки програми для для кластеризації та її візуалізації даних користувачів мережі Twitter перш за все необхідно було визначити мову програмування, основними критеріями вибору стали: наявність середовища для статичних обчислень, засобів реалізації кластерного аналізу даних та їх візуалізації.

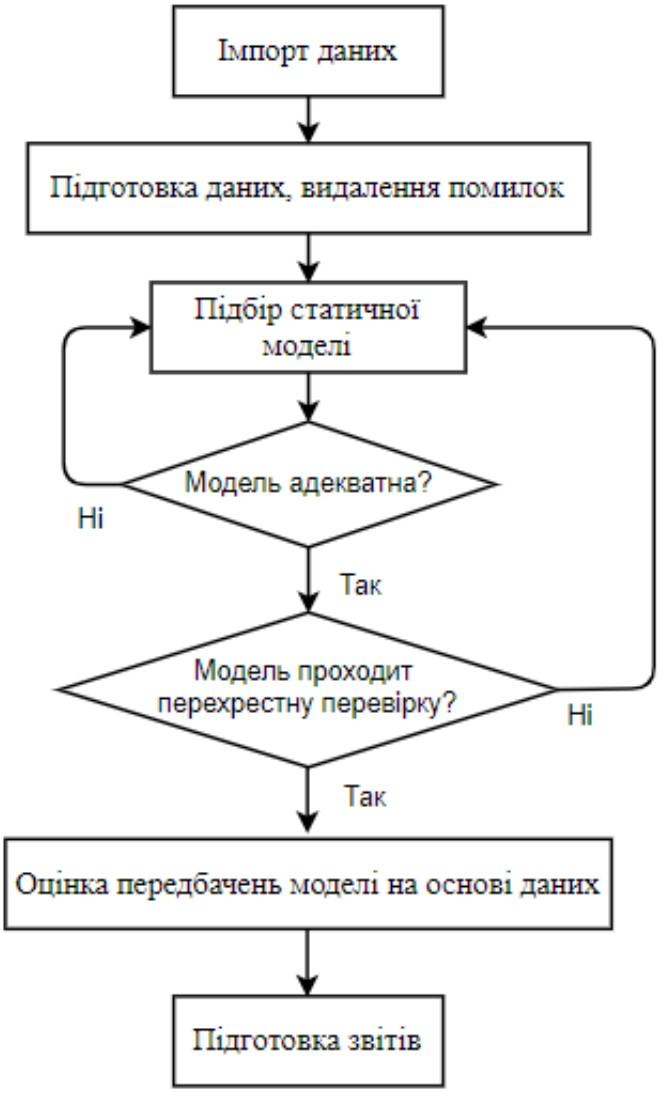

Рисунок 3.1 — Схема аналізу даних на мові програмування R

Таким чином була обрана мова R, підтримує об’єктно-орієнтовану, функціональну та масивову парадигми програмування, має динамічну типізацію, а також синтаксис, схожий на S з семантикою мови Scheme. Дана мова програмування підтримує як текстовий інтерфейс користувача, так і графічний. Також мова програмування має широкий вибір засобів проведення статичного аналізу, кластерного аналізу, аналізу часових серій [16].

Мова програмування R використовується у багатьох наукових сферах для аналізу даних та підготовки математичних моделей, включаючи біоінформатику та різні галузі медицини, економетрику, психометрію, машинне навчання та ін [17]. Приклад схеми аналізу даних, що може проводитись за допомогою мови програмування R можна побачити на рисунку 3.1.

Таким чином мова програмування R надасть можливість як проведення кластерного аналізу даних користувачів мережі Twitter, так і візуалізації результатів аналізу.

3.6 Опис графічного інтерфейсу користувача програми

При кожному запуску додатку користувачеві відображається основна форма програми (див. рис. 3.2). За допомогою кнопки “AddNew” можна викликати форму додавання нового користувача. Для того, аби переглянути інформацію певного користувача необхідно обрати мишею його Twitter нікнейм у списку ліворуч та натиснути кнопку “Show Full Info”. Щоб зберегти існуючих користувачів у файлову систему у вигляді excel файлу необхідно натиснути кнопку “Save”.

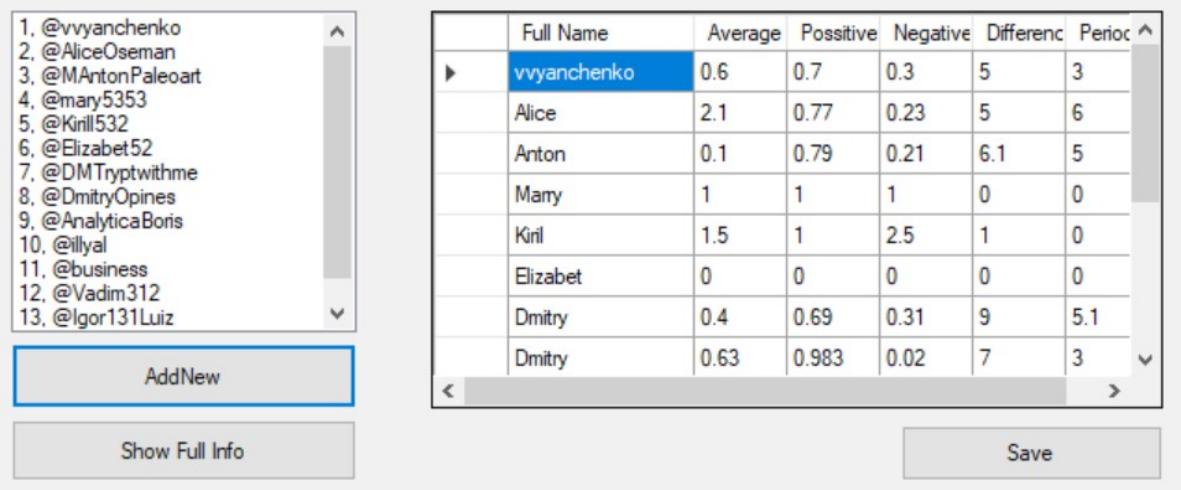

Рисунок 3.2 – Основне меню додатку

Також кнопка “AddNew” слугує відправною точкою у сценарії проведення сентимент-аналізу даних користувачів, адже дозволяє додавати дані необхідні для аналізу. По натисненню на цю кнопку користувачеві відображається вікно додавання даних користувача (див. рис. 3.3). Далі необхідно ввести у відповідні поля ім’я користувача, дату народження та Twitter нікнейм.

За допомогою використання C# бібліотеки HttpClient та Twitter Standard API програма знаходить ідентицікатор користувача та підставляє його у поле “Id”.

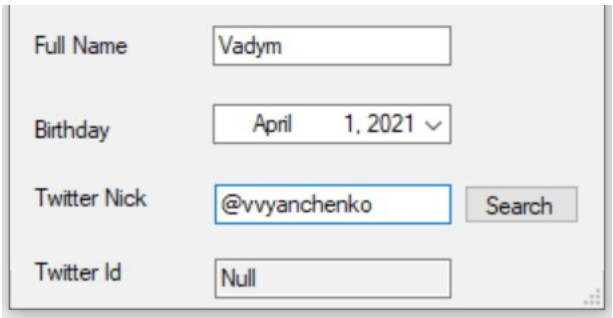

Рисунок 3.3 – Вікно додавання даних користувача

У випадку відсутності користувача з заданими параметрами програма відображає форму помилки (див. рис. 3.4).

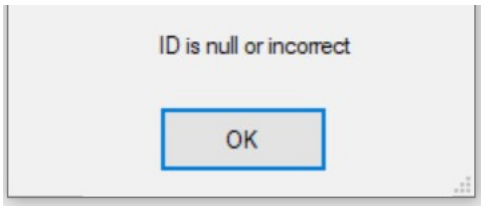

Рисунок 3.4 – Форма помилки

Після додавання інформації користувача додаток проводить аналіз настроїв на основі даних твітів. Для аналізу використовуються твіти за увесь період існування облікового запису даного користувача у мережі Twitter. Оскільки аналіз відбувається у фоновому режимі, дані користувачів на рисунку 3.2 оновлюються автоматично по завершенню обробки даних кожного користувача.

Для того, аби переглянути розгорнуту інформацію аналізу настроїв користувача необхідно відкрити відповідне вікно натиснувши кнопку “Show Full Info” (див. рис. 3.5).

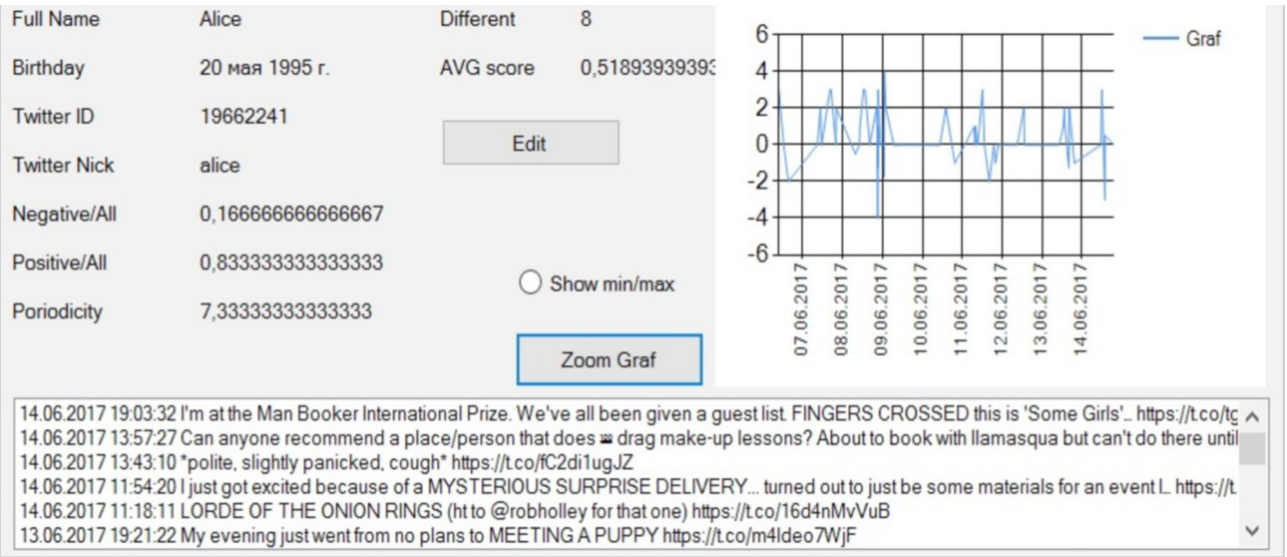

Рисунок 3.5 – Вікно з описом метрик психічного стану користувача

Дане вікно містить основну інформацію користувача, таку як нікнейм, ідентифікатор у мережі Twitter, дату народження та справжнє ім’я. Смислове значення метрик психічного стану буде описано у наступному абзаці. Тенденцію змін психічного настрою користувача можна відстежити за графіком зправа. Даний графік відображає залежність психічного стану користувача, вимірюваного за шкалою словника тональності, від дати публікації твіта. Також у окремій секції внизу вікна можна побачити основну інформацію про твіти користувача з початку реєстрації у мережі. Після того, як програма опрацює дані доданих користувачів, необхідно обрати кнопку “Save”, що знаходиться на основному вікні додатку (див. рис. 3.2) та дозволяє обрати шлях у провіднику файлової системи, куди необхідно зберегти дані аналізу настроїв.

Таким чином, після визначення імені файлу, результат аналізу тональності даних користувачів буде збережений у файловій системі у вигляді Microsoft Excel (див. рис. 3.6).

№	Full Name	Overall Average	Positive %	Negative %	Delta Min/Max	Frequency
1	Vadim	0.57	0.79	0.21	4	2.3
2	Alice	-0.1328	0.356	0.6438	8	7.1
3	Leo	-0.2941	0.7647	0.2352	5.5	4.25
4	Marry	1	1	0	0	1
5	Kate	2.5	1	0	1	1.5
6	Elizabeth	NaN	NaN	NaN	NaN	NaN
7	Dmitry	0.5438	0.7926	0.2074	9	4.68
8	Paul	0.8151	0.983	0.02	7	2.5
9	Boris	0.7096	0.8115	0.1846	7	4.64
10	Illya	-0.583	0.43	0.569	9	10.1
11	Alex	0.2043	0.798	0.2019	4	37
12	Yazhi	NaN	NaN	NaN	NaN	NaN
13	Igor	0.3	1	0	0	2
14	Daria	NaN	NaN	NaN	NaN	NaN

Рисунок 3.6 – Результати аналізу тональності даних користувачів

Як вже згадувалося, AFINN-111 являє собою набір пар ключ-значення, тобто словник пар слово до оцінки емоційного забарвлення цього слова. Через це в процесі сентимент-аналізу твіти користувача розбиваються на окремі слова та аналізуються послівно. Кожному слову присвоюється певна оцінка його емоційної забарвленості у відповідності до словника, далі для поточно-оброблюваного твіта знаходиться середня оцінка ступеню емоційної забарвленості усього твіта “Tweet Average”. Для того, аби зробити оцінку психічного стану користувача більш об’єктивною, програма знаходить значення метрики “Overall Average”, що підраховується як середнє арифметичне оцінок емоційної забарвленості кожного твіта.

Значення метрики “Delta Min/Max” обчислюється як різниця між максимальною та мінімальною оцінки твітів певного користувача. Дана метрика дозволяє оцінити психологічну стабільність користувача за певний проміжок часу.

За допомогою метрик “Negative %” та “Positive %” можна оцінити відсоток негативно-забарвлених та позитивно-забарвлених висловлювань. Сума значень метрик дорівнює 100%.

Метрика “Frequency” надає можливість оцінити наскільки часто користувач висловлюється у мережі Twitter, тобто це середнє число твітів за один день. Обчислюється даний показник як співвідношення загального числа твітів користувача до кількості днів з моменту реєстрації у мережі і до поточної дати.

На основі даних результату сентимент-аналізу твітів користувача далі проводиться кластерний аналіз з використанням засобів програмного середовища R (див. рис. 3.7).

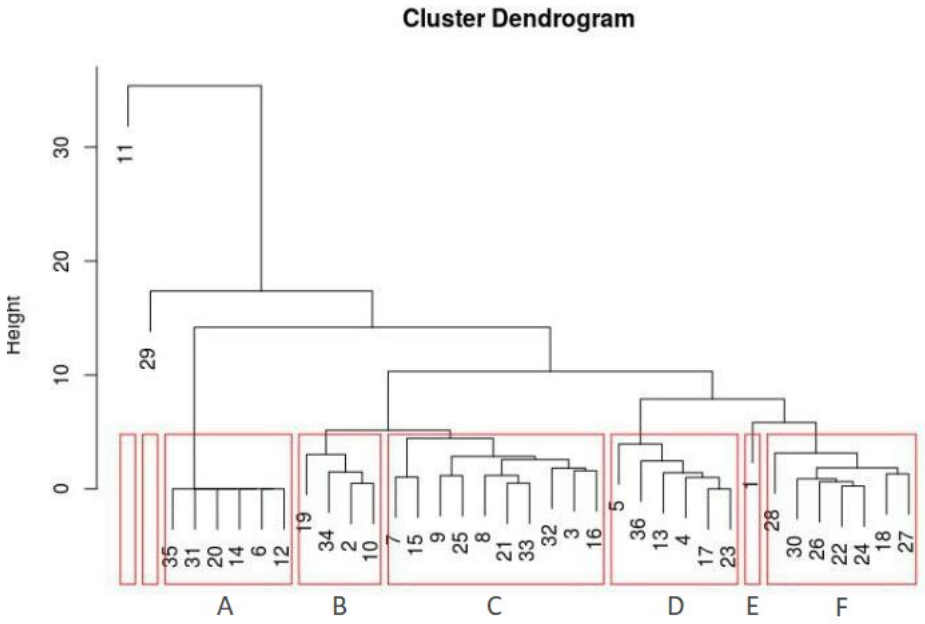

Рисунок 3.7 – Візуалізація дендрограми ієрархічної кластеризації

Проведення кластерного аналізу на основі методу ієрархічної кластеризації відбувалося у відповідності до метрик “Overall Average” (загальна оцінка психологічного стану користувача — середнє оцінок усіх твітів), “Delta Min/Max” (різниця максимального та мінімального значень “Overall Average”) та “Frequency” (середнє число твітів користувача протягом одного дня).

4 ОЦІНКА РЕЗУЛЬТАТІВ КЛАСТЕРНОГО АНАЛІЗУ

У результаті проведення кластерного аналізу методом ієрархічної кластеризації утворилися достатньо чітка структура кластерів дендрограми.

Першочергово слід розглянути на рисунку 3.7 користувачів кластеру “A”. Для користувачів цього кластеру значення усіх метрик текстового аналізу дорівнює “NaN”, тобто для даних акаунтів відсутні дані необхідні для проведення аналізу настроїв. Це може вказувати на той факт, що акаунти використовуються не за цільовим призначенням, а, до прикладу, є бот-акаунтами, що створені для штучного накручування вподобань чи репостів. Значення “Not a Number” стало причиною групування даних акаунтів у один кластер.

Розглядаючи кластер “B” можна знайти таку закономірність, що кожен користувач має високий показник метрики “Frequency”, тобто високе середнє число твітів на день. Проаналізувавши дану метрику для всіх користувачів, можна визначити, що значення більше 6 є завеликими. Також слід зазначити високі показники “Delta Min/Max” в цьому кластері — метрики, що опосередковано вказує на ступінь емоційної стабільності. Спираючись на зависокі числа згаданих метрик — середньої кількості твітів за день та дельти найбільшого та найменшого значення метрики “Overall Average” — можна припустити, що кластер “B” групує користувачів з психічно-нестабільним станом.

Розглядаючи кластер “C” можна стверджувати, що більшість показників метрик психічного стану знаходяться у межах норми, проте в цьому кластері також є помітні аномалії. До прикладу, зависокі показники метрики “Delta Min/Max” спостерігаються у користувача 7, а також у користувача 9, одночасно з нормальними показниками “Overall Average” та “Frequency”. Можна припустити, що дані користувачі переживали короткочасний стрес, який все ж

не погіршив загальний психічний стан та відносну небезпечність соціуму. Користувач 8 має високі показники позитиву у повідомленнях, згідно із аналізом настроїв 97% слів твітів користувача є позитивно забарвленими, незважаючи на завищену “Delta Min/Max” (подібно до користувачів 7 та 9) цей користувач психічно стабільний. Дивлячись на дещо негативний показник “Overall Average” у кристувача акаунту 3 можна стверджувати, що більшість твітів користувача є нейтрально або негативно забарвлені, проте беручи до уваги нормальні показними “Delta Min/Max” та “Frequency” можна припустити, що така поведінка є нормальною для нього і він соціально-безпечний і не чинить загрози іншим.

У результатах аналізу тональності на рисунку 3.6 можна побачити людей з показниками метрики “Overall Average” у діапазоні [0;1], у яких метрика “Frequency” показує відносно мале число твітів на день. Можна зробити припущення, що дані користувачі проявляють низьку активність у мережі. Такі користувачі були згруповані кластерним аналізом у кластер “D”.

Якщо переглянути показники психологічного стану користувачів кластерів “E” та “F”, то можна переконатись, що вони є психічно стабільними.

Таким чином, у результаті проведення кластерного аналізу методом ієрархічної кластеризації, були сформовані наступні кластери та їх характеристики:

- кластер “A” — користувачі з “пустими” акаунтами, де немає жодного твіта;
- користувачі 11 та 29 — соціальні медіа-акаунти (спільноти, інфлюенсери або блогери тощо);
- кластер “B” — психічно-нестабільні користувачі;
- кластер “C” — загалом психічно-стабільні користувачі, з присутніми аномаліями певних метрик психічного стану;
- кластер “D” — психічно-стабільні користувачі з низькою активністю у мережі та невеликим середнім числом твітів на день;

– кластери “E” та “F” — психічно-стабільні користувачі.

Таким чином, оцінюючи результати проведення кластерного аналізу методом ієрархічної кластеризації за ознаками “Overall Average” (загальна оцінка психологічного стану користувача на основі твітів — середнє арифметичне відповідних оцінок усіх твітів), “Delta Min/Max” (різниця максимального та мінімального значень “Overall Average”) та “Frequency” (середнє число твітів користувача протягом одного дня) можна зробити висновок про успішність процедури, оскільки вдалося досягти розбиття даних щодо психічного стану користувачів на чітко сгруповані кластери які можна логічно описати та охарактеризувати [18].

Єдиним недоліком обраних ознак є наявність на дендрограмі соціальних “медіа-акаунтів”, а саме акаунтів 11 та 29, що знаходяться у різних кластерах, адже в них значно відрізняються показники метрики “Frequency”. Проблему групування таких користувачів можна подолати шляхом введення додаткового критерію кластеризації, такого як, наприклад, “Followers” — число підписників обраного акаунту.

5 АНАЛІЗ МОЖЛИВИХ ЗАСТОСУВАНЬ

Як тільки людина входить до мережі Інтернет, вона починає залишати за собою свої «цифрові сліди» за якими сайти, які людина відвідала, будуть відстежувати всі її дії онлайн. Це абсолютно законні дії. Ці данні включають місцезнаходження людини, її IP-адресу, а також інформацію про операційну систему, чи це стаціонарна або мобільна версія операційної системи, модель процесора і навіть рівень зарядження акумулятора для ноутбуків, планшетів або телефонів.

Але нас цікавлять дані користувачів соціальних мереж (Twitter, Facebook, Instagram), які люди залишають (і навіть розповсюджують) самостійно – у вигляді постів та коментарів, тобто набір тексту.

Мета даної кваліфікаційної роботи – визначити по цьому набору тексту для кожного користувача соціальної мережі психічний та функціональний стан людини, а у перспективі потенційно і його реальні цілі та мотиви. Це дає можливість виявити наявність визначених психологічних проблем, допоможе спрогнозувати поведінку людини в будь-якій ситуації, визначити сумісність, позитивні та негативні якості, зрозуміти настрій в колективі і не тільки. Звідки залежить і вірогідна галузь застосування: виявлення терористів (для СБУ), вбивць, педофілів та соціально небезпечних осіб (для поліції), психічно хворих людей (для соціальних служб) та і просто під час прийняття на роботу – визначити чи здатна людина виконувати свої обов’язки.

Аналіз тональності – це визначення загально емоційного забарвлення поста в соціальної мережі певної людини. Кожне слово поста має своє емоційне забарвлення в діапазоні від -5 до +5 балів, сумуючи це забарвлення всіх слів ми отримуємо загальне емоційне забарвлення поста, як емоційно він забарвлений позитивно чи негативно. Аналізуючи набір таких постів людини ми зможемо зробити припущення про її психічний стан.

Тобто начальник зможе відправити співробітника, який знаходиться у стані стресу, у відпустку. Людині, яка знаходиться у пригніченому стані – надати допомогу кваліфікованого лікаря психолога чи психіатра. А людині яка відчуває наснагу – довірити складне завдання.

Подібний аналіз був би у нагоді у процесі найму людей до державних органів. Таким чином можливо було б визначити ступінь психічної врівноваженості та зробили припущення щодо того, чи здатен кандидат виконувати прямі обов'язки.

Не зважаючи на досить велику кількість можливих корисних застосувань методу, необхідно брати до уваги особливості інформації, що піддається аналізу, адже порушення особистого простору людини може призвести до серйозних соціальних незадоволень та збурень. Адже приватне життя – це потреба у власному просторі. Згідно цивільному кодексу України: «Фізична особа має право вільно збирати, зберігати, використовувати і поширювати інформацію. Збирання, зберігання, використання і поширення інформації про особисте життя фізичної особи без її згоди не допускаються, крім випадків, визначених законом, і лише в інтересах національної безпеки, економічного добробуту та прав людини.» [19].

Проте, аналіз психічного стану людини може бути частково неефективним та неповним у тому випадку, якщо аналіз настроїв базується лише на тих даних, що людина показує сама, тобто повністю публічних даних.

Для початку за основу джерел даних, на яких можна застосовувати методики аналізу психічного та фізіологічного стану людини, можна взяти публічні ресурси, такі як різноманітні форуми. Питання використання аналізу настроїв для особистих повідомлень є дуже суперечливим. Проте, гіпотетичне застосування такого підходу стало би у нагоді спецслужбам в інтересах національної безпеки. Тим паче якщо провести групування усіх акаунтів певної людини у єдиній базі.

Не дивлячись на те, що на основі сентимент-аналізу можна робити припущення про наміри людини, постає потреба у тренованій нейромережі, що базуючись на спеціальних словниках тональності змогла б визначати вірогідність певних дій користувача, провести так званий “аналіз намірів” або “intent analysis”. У такому випадку, ймовірно, було б доречно використати вірогіднісний метод “c-means” для визначення ймовірності виконання певної дії та детально підібрати критерії кластеризації. Проте, таких підхід буде малоефективним, якщо не аналізувати персональні повідомлення користувача.

Застосування аналізу настроїв може бути більш місцевим, до прикладу у приватному житті. Наприклад, якщо людина схильна до гніву та агресії і схоче відстежувати свій психологічний стан для своєчасної надання допомоги. Або сім’ї в яких є людина, яка схильна до депресії та суїциду. Якщо така людина, розуміючи ризики надає своїй сім’ї доступ та дозвіл для аналізу настроїв, то це може врятувати їй життя.

Також батьки мали б змогу контролювати психічний стан своєї дитини. До психічних розладів у дітей можуть відноситися незначні проблеми, які легко корегуються, а також більш серйозні процеси, які ведуть до розладу психіки дитини. Часто психічні відхилення залишаються непомітними у зв’язку з неуважністю батьків або їх браком часу. Насправді, чим раніше буде виявлена проблема у розвитку дитини, тим більш ефективні будуть заходи для її корекції. За використання даної методики дитячі психологи змогли б детальніше проаналізувати причину психологічних проблем пацієнтів і підібрати найбільш ефективні підходи лікування. Однак на заваду стає питання етичності подібного стеження, навіть якщо це власні діти.

Також для батьків дуже важливо розуміти психологічний клімат у групі друзів дитини. Для цього ми можемо використати метод кластерного аналізу для групування у психічно стабільні, психічно нестабільні та психічно критичні кластери. Батьки змогли б обмежети спілкування своєї дитини з останньою групою осіб.

ВИСНОВКИ

У результаті досліджень виконаних у ході кваліфікаційної роботи був проведений аналіз предметної галузі та аналіз літератури за обраною тематикою, що дозволило проаналізувати актуальні тенденції та дослідження у напрямку оцінки думок людини за її текстовими висловлюваннями та зробити обгрунтовані рішення щодо обрання об’єкту проектування. Також був проведений аналіз існуючих методів оцінки психічного та функціонального стану людини на базі її думок висловлених у текстовому форматі, таким чином було вирішено використати синтез аналізу тональності тексту та кластерного аналізу сентимент-даних.

З метою обрання найбільш доцільного за обраною тематикою методу кластерного аналізу був проведеним системний аналіз методів кластерного аналізу. Першим кроком були обрані методи кластерного аналізу для дослідження (за фактором популярності та приналежності до різних кластерних моделей) та критерії їх оцінки (універсальність алгоритму, складність, кількість необхідних даних, точність кластерного аналізу, необхідні об’єми обчислень). Згодом у рамках ієрархічного аналізу на основі обраних критеріїв та методу парних порівнянь був обраний найбільш ефективний метод кластерного аналізу, яким виявився метод ієрархічної кластеризації, в основному за рахунок точності та меншого обсягу обчислень.

На основі методу ієрархічного аналізу був проведений аналіз проблем кластеризації, у ході якого були обрані категорії задоволення та відповідні властивості що у комбінації з методом парних порівнянь дозволило визначити критично важливі критерії системи, а саме зниження похибки прогнозу, зниження складності програмної реалізації, а також підвищення точності результатів.

Також були окремо більш детально розглянуті особливості, переваги та недоліки з точки зору алгоритму методів кластерного аналізу. Метод ієрархічної кластеризації, у противагу альтернативам, не потребує ручного визначення кількості кластерів користувачем, що є значною перевагою, тому що досить складно визначити самостійно кількість кластерів що будуть сформовані на основі даних аналізу тональності твітів. Також перевагою методу є відображення схожості кластерів завдяки висоті ребера між цими кластерами, що надає змогу виокремити отримані кластери та покращує точність припущень щодо особливостей отриманих кластерів.

З метою підготовки даних для кластерного аналізу був розроблений програмний додаток на мові програмування C#, що надає зручний інтерфейс користувача для завантаження текстових даних користувачів соціальної мережі Twitter, отриманих за допомогою Twitter API, їх перетворення у загальне нормалізоване представлення та відображення статистики. Також був розроблений додаток на мові програмування R, що дозволяє провести кластерний аналіз на основі отриманих попередньою програмою нормалізованих даних користувачів Twitter.

У результаті кластерного аналізу були сформовані чіткі кластери користувачів, згруповані за обраними ознаками (середня оцінка тональності твітів, дельта граничних оцінок та частота твітів), які вдалося проаналізувати у відповідності до оцінок, отриманих на етапі аналізу тональності, та зробити припущення щодо психічного та функціонального стану користувачів, що входять у отримані кластери. Вдалося отримати умовні кластери психічно-стабільних користувачів, психічно-стабільних користувачів з присутніми аномаліями певних метрик психічного стану та психічно-нестабільних користувачів. Такий аналіз дозволив переконатися у тому, що обраний метод кластерного аналізу виявився найбільш ефективним для поставленої задачі.

Варто зазначити, що різні люди по-різному висловлюють свої думки та емоції. Чорний гумор, сарказм та іронія, протиставлення понять та

багатозначність деяких слів — все це значно ускладнює задачу аналізу тональності тексту та зменшує точність такого аналізу. Тому на основі лише сентимент-аналізу не можна точно стверджувати, що певна особа несе загрозу того чи іншого характеру. Проте, спільне використання аналізу тональності тексту та методів кластеризації дозволяє робити певні припущення про психічний та функціональний стан людини, помітити певні аномалії у поведінці людей. Також слід звернути увагу на те, що аналіз отриманих у ході дослідження кластерів є людським аналізом, тобто суб'єктивним. Це є як певним недоліком обраного у даній кваліфікаційній роботі методу аналізу кластерів, так і простором для вдосконалення дослідження у майбутньому. Дану проблему можна вирішити за допомогою використання алгоритмів машинного навчання, так званих класифікаторів, що могли б робити припущення щодо вірогідних дій певної особи чи групи осіб у кластері та класифікувати такі дії за рівнем загрози того чи іншого характеру.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Informativity of Association Rules from the Viewpoint of Information Theory / D. Sitnikov, O. Titova, S. Minukhin, A. Kovalenko, S. Titov // 2018 International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 - Proceedings, 2019, c. 595-598.

2. Cesarano C, Dorr B, Picariello A, Reforgiato D, Sagoff A, Subrahmanian VS. Oasys: an opinion analysis system. In: AAAI spring symposium on computational approaches to Analyzing Weblogs, 2004.

3. Chaovalit P, Zhou L. Movie review mining: a comparison between supervised and unsupervised classification approaches. In: Proceedings of the 38th Hawaii international conference on system sciences, IEEE Computer Society, 2005.

4. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Conference on empirical methods in natural language processing (EMNLP). Philadelphia, Pennsylvania, USA, 2002, p. 79.

5. Turney PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: 40th annual meeting of the Association for Computational Linguistics (ACL), Philadelphia, Pennsylvania, USA, 2002, p. 417.

6. Wiebe JM. Learning subjective adjectives from corpora. In: Conference on artificial intelligence, Menlo Park, CA. AAAI Press 2000, pp. 735–741.

7. Turney P. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: European conference on machine learning. Berlin: Springer, 2001, p. 491.

8. Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004, p. 271.

9. Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting of the Association for Computational Linguistics, 2005, pp. 115–124.

10. Das SR, Chen MY. Yahoo! for Amazon: sentiment extraction from small talk on the web. Management Science 2007; 53: 1375–1388.

11. Wenhong Tian, Yong Zhao. Optimized Cloud Resource Management and Scheduling: Theories and Practices. – Morgan Kaufman, 2014. – 284 p.

12. Ирина Чубакова. Data Mining. Методы классификации и прогнозирования. URL: http://www.intuit.ru/studies/courses/6/6/lecture/174 (дата звернення: 05.04.2021).

13. Ulrich Meyer, Peter Sanders. Algorithms for Memory Hierarchies: Advanced Lectures. – Springer Science & Business Media, 2003. – 428 p.

14. Seetha Hari, Murty Narasimha, B.K. Tripathy. Modern Technologies for Big Data Classification and Clustering. – IGI Global, 2015 – 360 p.

15. Cliff Goddard. Semantic Analysis: A Practical Introduction. – OUP Oxford, 1998. – 428 p.

16. Рассел Метью, Классен Михайло. Data mining. Извлечение информации из Facebook, Twitter, LinkedIn, Instagram, GitHub. – Питер, 2020. – 464 с.

17. Hadley Wickham, Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. – O'Reilly Media, 2017. – 520 p.

18. Бокс Дж. Дженкинс Г. Анализ временных рядов: прогноз и управление. – Мир, 1974. – 405 с.

19. Джоел Грас. Data Science. Наука о данных с нуля. – БХВ-Петербург, 2016. – 336 с.

20. Конституція України – Розділ II – Стаття 32. URL: https://www.president.gov.ua/ua/documents/constitution/konstituciya-ukrayini-rozdil-ii (дата звернення: 13.04.2021).