МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА

РОБОТА 15.03 — КМР. 1940–“С” 2022.12.30.

011 ПЗ СЕМЕНКО АЛІНИ АНАТОЛІЇВНИ

2023 р.

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

УДК 004.77

«ПОГОДЖЕНО»	**«ДОПУСКАЄТЬСЯ ДО ЗАХИСТУ»**
Декан факультету	Завідувач кафедри комп'ютерних
інформаційних технологій	наук
Глазунова О.Г., д.пед.н., професор	Голуб Б.Л., к.тех.н., доцент
_____	_____
« » _____2023 р	« » _____2023 р

МАГІСТЕРСЬКА РОБОТА

На тему: Технології десимінації та аналізу використання цифрового контенту у мережі інтернет.

Спеціальність 122 Комп'ютерні науки
(шифр і назва)

Освітньо-професійна програма Інформаційні управляючі системи та технології
(назва)

Робота на здобуття кваліфікації магістра

Керівник магістерської роботи
доктор пед.наук, професор / Глазунова О.Г. /
(вчене звання і ступінь) (підпис) (ПІБ)

Виконала _____ / Семенко А.А. /
(підпис) (ПІБ студента)

КИЇВ – 2023

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ЗАТВЕРДЖУЮ
Завідувач кафедри
комп'ютерних наук
(назва кафедри)
к.тех.н., доцент Голуб Б.Л.
(вчене звання і ступінь) (підпис) (ініціали і прізвище)

«_____» _____ 2023 р.

З А В Д А Н Н Я

ДО ВИКОНАННЯ МАГІСТЕРСЬКОЇ РОБОТИ СТУДЕНТУ

Семенко Аліни Анатоліївни
(прізвище,ім'я,по-батькові)
Спеціальність 122 Компютерні науки
(шифр і назва)
Освітньо-професійна програма Інформаційні управляючі системи та технології
1. Тема магістерської роботи: Технології десимінації та аналізу використання цифрового контенту у мережі інтернет

затверджено наказом ректора НУБіП від « 30 » грудня 2022 р. № 1940-С
Термін подання завершеної роботи на кафедру 5 листопада 2023 року
(рік, місяць, число)

3. Вихідні дані до магістерської роботи:
Сторінки соціальніих мережах Facebook, Twitter, Instagram та вебсайт Ecotwins. Основні критерії: загальна кількість вподобань, тип запису, категорія, місяць публікації, день тижня публікації, час, чи є рекламною проплатою, загальний обсяг публікацій, загальна кількість переглядів, кількість коментарів, кількість вподобань, кількість поширень, кількість «підписаних» осіб серед переглядів/вподобань, обсяг залучення нової аудиторії, загальна кількість взаємодій із записом та інші.

4. Перелік питань, що підлягають дослідженню:
1. Дослідити методи збору даних зі сторінок у соціальних мережах: Facebook, Twitter, Instagram та на вебсайті Ecotwins для передачі їх в систему аналізу.
2. Спроектувати та розробити систему аналізу.
3. Дослідити та обґрунтувати мету і техніку аналізу даних
4. Дослідити параметри, за якими можна сформулювати рекомендації адміністраторам сторінок Facebook, Twitter, Instagram та на вебсайті Ecotwins

5. Перелік графічного матеріалу (за потребами): постер

Дата видачі завдання "17'' грудня 2022 р.
Керівник магістерської роботи Глазунова О.Г.
(підпис) (прізвище та ініціали)
Завдання прийняв до виконання Семенко А.А.

(підпис) (прізвище та ініціали студента)

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

OLAP – online analytical processing, аналітична обробка в реальному часі.

KPI – Key Performance Indicators, ключовий показник ефективності (успішності).

BI – Business intelligence.

СД – сховище даних.

БД – база даних.

ІС – інформаційна система.

MS – Microsoft.

SQL – structured query language, мова структурованих запитів.

OLE BD – Object Linking and Embedding, Database.

ЦА – цільова аудиторія.

ВСТУП

У 2023 році, у 21 столітті ми не уявляємо свого життя свого життя без соціальних мереж, таких як: Facebook, Instagram, Twitter. Ми обмінюємося фотографіями та відео зі свого життя, слідкуємо за актуальними новинами, спілкуємося навіть на великій відстані та поглиблюємо свої знання в обраних сферах інтересів.

Серед 78% відсотків українського населення,котрі котристуються соцмережами, кожен 3-й є адміністратором хоча б однієї сторінки, хоча б в одній соціальній мережі. Але не всі користувачі використовують максимум можливостей, які надають соціальні мережі, оскільки багато з них не мають необхідної інформації або навичок. Це може призвести до обмеженого розвитку або навіть припинення існування сторінок через недосвідчених адміністраторів та контент-мейкерів.

Тому, **метою дослідження** є обгрунтування ефективності технологій десимінації цифрового контенту та аналізу його використання у мережі інтернет для підвищення активності та просування сторінок у соціальних мережах: Twitter, Facebook, Instagram та на вебсайті проекту.

Об'єктами дослідження відкриті сторінки проекту ECOTWINS є сторінки у соціальних мережах: Twitter(X): «ECOTWINS-Project about researching», Facebook: «ECOTWINS-Project about researching», Instagram: «ECOTWINS-Project about researching» та вебсайт «Ecotwins», веб-сайт: «ECOTWINS-Project about researching».

Предмет дослідження: система поширення та аналізу активності користувачів сторінок проекту у соціальних мережах: Twitter(X), Facebook, Instagram та на вебсайті проекту Ecotwins.

У цьому дослідженні ми створюємо систему, яка аналізує відкриті сторінки користувачів в соціальних мережах: Facebook, Twitter, Instagram та

вебсайт; обраховує рівень медіа активності відвідувачів, збирає кількісні та критеріальні дані, оброблює та робить висновки - рекомендації.

У якості даних, що потребують аналізу використовуються відкриті дані сторінок у соціальних мережах Facebook, Twitter(X), Instagram: тип запису, категорія, місяць публікації, день тижня публікації, час, загальний обсяг публікацій, загальна кількість переглядів, кількість коментарів, кількість вподобань, кількість поширень, кількість «підписаних» осіб серед переглядів/вподобань, обсяг залучення нової аудиторії, загальна кількість взаємодій із записом та інше. Для аналізу активності відвідувачів на вебсайті також використовуються відкриті дані:джерела трафіку, поведінка користувача, відстеження конверсій та інше.

За результатами дослідження буде сформована інформаційна довідка для адміністраторів та контент-мейкерів сторінок в соціальних мережах та на вебсайті проекту.

У висновку, **завданням** дослідження є: розробка системи, яка дасть можливість визначити ефективність застосування технологій десимінації та аналізу до цифрового контенту у мережі Інтернет та соціальних мережах.

Актуальність дослідження визначається потребою знайти та застосувати методи для підвищення активності користувачів у медіа та залучення цільової аудиторії до сторінок проєкту Ecotwins у соціальних мережах Facebook, Instagram, Twitter(X) і на веб-сайті проєкту.

Проблема актуальна через декілька причин:

1. **Залучення цільової аудиторії:** Привертання цільової аудиторії є стратегічно важливим завданням для будь-якого проекту. Дослідження допоможе виявити, якими засобами та методами, можна залучити цільову аудиторію.
2. **Збільшення медіа-активності:** Спонукання користувачів до активної участі у соцмережах та на веб-сайті є важливим для формування активної спільноти навколо проекту. Дослідження допоможе визначити, які дії та контент стимулюють користувачів до взаємодії.

3. **Оцінка результативності стратегій:** Дослідження допоможе визначити, наскільки ефективними є поточні медіа-стратегії та як їх можна вдосконалити, шляхом виявлення недоліків і пошуку нових можливостей для розвитку.
4. **Моніторинг результатів:** Слідкування і аналіз результатів мають вирішальне значення для того, щоб в реальному часі коригувати стратегію і досягати заданих цілей.

Це дослідження сприятиме покращенню взаємодії проекту Ecotwins з аудиторією та досягненню більшого успіху в соціальних мережах та на веб-сайті.

Для пошуку таких шляхів, адміністратор та контент-мейкер повинні провести аналіз ключових аспектів управління сторінками в різних соцмережах, визначити особливості поведінки користувачів в них та на вебсайті та визначити основні бізнес-фактори. Аналіз діяльності користувачів на сторінках різних соцмереж та на вебсайті дозволив визначитись у таких питаннях:

- В який час найкраще публікувати записи? (Час, день тижня)
- Чи впливають наявність геолокації та відміток людей на загальну кількість перегляду публікації?
- В якій соціальній мережі найкраща активність користувачів?
- Який контент найбільш сприятливий для відвідувачів?
- Які дії найбільше роблять користувачі при відвідуванні вебсайту?

З'являється необхідність створити систему для збирання важливих даних, їх подальшої обробки та систематизації, проведення складного аналізу кількісних та якісних даних на сторінках у соціальних мережах та на веб-сайті, а також підготовки звітів. На основі цих звітів будуть розроблені рекомендації для підвищення технічних та адміністративних можливостей цих сторінок.

Апробація постановки проблеми та аналізу рішень дослідження відбулася в рамках V Всеукраїнської науково-практичної інтернет-конференції студентів та аспірантів «Теоретичні та прикладні аспекти розробки комп’ютерних систем 2023» 26 квітня 2023 (додаток А).

Результати дослідження та доцільність впровадження системи було апробовано в рамках XIV Міжнародної науково-практичної конференції молодих вчених «Інформаційні технології: економіка, техніка, освіта 2023» 26-27 жовтня 2023 року м. Київ, Україна (додаток Б) та розроблено постер.

Пояснювальна записка включає: вступ, три розділи, висновки, джерела та додатки. У вступній частині роботи розглядаються аспекти, що характеризують область, що вивчається, та коротко описується поточний рівень вирішення поставленої проблеми та взаємозв'язок із іншими проведеними дослідженнями. Тут визначається мета дослідження, об'єкт, предмет та завдання, також вказується схвалення проведеної роботи і надається структура пояснювальної записки.

Перший розділ включає аналіз постановки задачі та предметної області, розкриття постановки завдання, подання діаграми прецедентів та опис архітектури системи.

Другий розділ присвячений аналізу методів і технологій поширення контенту. Він також висвітлює зміст та сутність системи, її інформаційне забезпечення та технічну компоненту, включаючи вузли системи, які постачають дані до сховища, загальні концепції щодо OLAP-технологій та структуру сховища даних.

У третьому розділі розглядається створення системи поширення, включаючи механізми збору, обробки і передачі даних, опис BI та створення проекту служби SSAS (побудова розгорнутого куба), розробку звітів та розрахунок KPI. Крім того, цей розділ містить рекомендації для адміністраторів сторінок в соціальних мережах та контент-мейкерів вебсайту проекту Ecotwins, щоб краще керувати сторінками в різних соціальних мережах.

У розділі висновків наведена оцінка отриманих результатів дослідження та визначені можливості використання отриманих даних.

До пояснювальної записки також включено постер, на якому показані основні аспекти проведеної роботи.

1 АНАЛІЗ ТА ПРОЕКТУВАННЯ

Мета аналізу і проектування полягає в створенні інструкцій, які чітко визначають завдання та процес виконання роботи відповідно до вимог.

У цьому контексті можна виокремити такі завдання:

1. Розробка детальної архітектури розподіленої програмної системи.
2. Трансформація вимог у проектну модель розроблюваної системи.
3. Адаптація проекту системи для реалізації з метою підвищення продуктивності розробки.
4. Вибір механізмів реалізації та встановлення обмежень на їхню реалізацію.
5. Розробка компонентної структури системи.
6. Розподіл компонентів між вузлами. Аналіз надає чітке розуміння завдання та сформульованої проблеми, яку необхідно вирішити під час роботи.
7. На етапі проектування відбувається уточнення результатів аналізу з метою оптимізації роботи з урахуванням обмежень, таких як нефункціональні вимоги та середовище реалізації.

1.1 Постановка завдання

На етапі постановки завдання слід чітко визначити цілі та шляхи їх досягнення. Тут формуються конкретні завдання, включаючи опис вхідних та вихідних даних.

Вхідна інформація щодо завдання - це дані, які подаються на вхід системи та використовуються для вирішення цього завдання.

Вихідна інформація може бути представлена у формі документів, зображень на екрані монітора, даних у базі даних або вихідного сигналу пристрою управління.

Головним завданням цього дослідження є створення інформаційного листу для адміністраторів сторінок у соціальних мережах та контент-мейкерів

вебсайту міжнародного дослідницького проекту Ecotwins. Однак, оскільки проект існує всього лише рік, дані, які потребують подальшого аналізу, мають приблизний характер і визначають цілі, яких нам необхідно досягти для успішного розвитку нашого проекту.

Джерелами інформації виступили: веб-сайт і три сторінки в різних соціальних мережах проекту Ecotwins, а саме:

- Twitter: «ECOTWINS-Project about researching»;
- Facebook: «ECOTWINS-Project about researching»;
- Instagram: «ECOTWINS-Project about researching».

Для аналізу була використана відкрита інформація, доступна на сторінках у соціальних мережах. Серед головних параметрів, що підлягали аналізу, включалися такі аспекти як вид публікацій, їх категорія, дата (місяць та день тижня) і час публікацій, кількість опублікованих записів, кількість переглядів, кількість коментарів, кількість лайків, кількість поділень, кількість підписників серед переглядачів та лайків, приваблення нової аудиторії, загальна кількість взаємодій із публікаціями та інші показники.

Під час збору даних мало місце дотримання уставів Закону про захист даних та політики використання даних компанії Meta(в компанію вхоять соціальні мережі Facebook, Instagram), компанії Twitter(X) та вебсайту проекту Ecotwins [25]. У ході дослідженя використовувалися персональні дані користувачів соціальних мереж Фейсбук, Інстаграм та Твітер(X) – особиста інформація, яка може бути використана для визначення або ідентифікації конкретної особи, така як ім'я, адреса, контактні дані (номер телефону, адреса електронної пошти) і т. д. Виконавець дослідження дотримувався політики конфіденційності компанії та дотримувався вимог законодавства України, включаючи закони про захист персональних даних, інформацію, рекламу, телекомунікації, підприємництво, а також закони про захист інформації в інформаційно-телекомунікаційних системах. Додатково, були застосовані вимоги, визначені в Національному документаційному технічному стандарті

НД ТЗІ 2.5-010-03 "Вимоги до захисту інформації на веб-сторінці від несанкціонованого доступу" [13].

Збір даних у соцмережах Facebook та Instagram та проводився за допомогою додатку Facebook Business Suite. Facebook Business Suite надає розширені інструменти аналітики для бізнесів, які допомагають відстежувати та аналізувати результати їхньої присутності на соціальних мережах, зокрема на Facebook та Instagram [14].

В додатку Business Suite власники сторінок мають можливість отримувати сповіщення, повідомлення та повідомлення від Facebook, Instagram і Messenger. Більше того, ця програма дозволяє одночасно управляти корпоративними обліковими записами на трьох різних платформах, створювати та планувати публікації і навіть налаштовувати рекламні кампанії. Приєднання сторінок до додатку "Facebook Business Suite" надає можливість досвідченим програмістам завантажувати необхідні дані у різних форматах або навіть проводити аналіз безпосередньо всередині програми.

Дані з сайту Ecotwins було зібрано за допомогою ресурсу Google Analytics- це безкоштовний інструмент для аналізу веб-сайтів, створений компанією Google. Google Analytics допомагає веб-мастерам і бізнес-власникам зрозуміти, як користувачі взаємодіють з їх веб-контентом, щоб вони могли приймати кращі рішення для поліпшення та оптимізації свого веб-продукту [16].

Для аналізу поширення твітів(постів) в Twitter можна використовувати різні методи та інструменти для збору та обробки даних: Twitter API(Application Programming Interface)- набір інструментів , створених Twitter для розробників, які дозволяють отримувати доступ до функціоналу та даних соціальної мережі Twitter з метою створення різних програм, додатків та інших рішень, які інтегруються з цією платформою [24].

Всі ці платформи є безкоштовними і нададуть найточніший аналіз даних з соцмереж та сайту (Додаток А).

Існують численні сервіси, які можуть виконувати схожі функції, незалежно від загальних засобів доступу до даних. Проте важливо зазначити, що не всі ці сервіси є безкоштовними, і не всі з них надають можливість вивантажувати дані з мережі.

Серед сервісів можливих для використання під час дослідження є:

- Similar Web – цей інструмент аналітики допомагає дослідити світові ринкові тенденції, проаналізувати соцмережі ваших конкурентів та зрозуміти поведінку аудиторії. Навіть за допомогою безплатної версії можна визначити, які канали у вашій ніші дають більшу частину трафіку, та створити порівняльний аналіз із конкурентами за обраними показниками. Вартість послуг від 250$ на місяць, також є безкоштовна версія на місяць;
- Klear – це комплексний інструмент, який допомагає знайти інфлюенсерів та лідерів думок, що відповідатимуть вашій цільовій аудиторії та цінностям вашого бренду. Для цього програма аналізує десятки показників, як-от тематику профілю інфлюенсера, інформацію про підписників, теми дописів. У результаті система ставить оцінку від 0 до 100, яка кількісно визначає вплив інфлюенсера на користувачів. А технологія штучного інтелекту FakeSpot допомагає виявити у цих інфлюенсерів накручених підписників та будь-які штучні взаємодії із профілем, що створюють боти. Вартість послуг від 250$ на місяць, також є безкоштовна версія на місяць;
- Social mention – репутаційний інструмент, який допомагає розширити уявлення про те, як клієнти взаємодіють з вашим брендом та як працює ваша стратегія в соціальних мережах. Проста в користуванні безкоштовна програма, відстежує згадки про бренд/компанію в блогах, новинах, відео, статтях, відгуках та скаргах. Програма збирає дані з понад 100 соціальних мереж:Twitter, Facebook, FriendFeed, YouTube, Digg, Google.

- Majestic API - система веб-аналітики, однією з основних функцій яких є аналіз посилань. З її допомогою з'ясувати, які посилання ведуть на ваш сайт, з'ясувати, звідки приходять користувачі і по яких посиланнях вони клацають найчастіше. Оцінюється і контекст, в якому знаходяться посилання. Також є можливість відслідковувати пов'язані сайти і сайти, які посилаються не тільки на вас, але і на конкурентів [18, 19].

Після дослідження багатьох сервісів та методів аналізу даних, виявлено, що саме Business Suite, Twitter API та Google Analytics виявилися універсальними інструментами для даного дослідження.Необхідною вимогою до виконання роботи є концентрація зібраної інформації в базі даних.

База даних (БД) - це система, створена для структурованого зберігання, редагування та обробки об'ємних наборів взаємопов'язаної інформації. Вона включає у себе набір даних, які організовані відповідно до певної концепції, що визначає характеристики цих даних та їх взаємозв'язки.

Для розробки бази даних використовувалася SQL Server, яка є однією з найбільш поширених систем управління базами даних в сучасності. Ця система призначена для опрацювання значних обсягів інформації. Однією із ключових переваг цієї системи є її доступність, оскільки використання цього програмного продукту є безкоштовним.

На основі бази даних було створено систему для зберігання, управління та обробки інформації, відому як сховище даних. Ця система дозволяє зберігати різноманітні дані та надає доступ до них через різні методи, для здійснення пошуку, оновлення та аналізу інформації. Зазвичай дані в сховищі доступні лише для читання.

Важливо зазначити, що проектування баз даних і систем обробки даних не є самоціллю. Оперативна база даних, яка використовує технологію OLTP (Online Transaction Processing), використовується в автоматизованих системах обробки інформації (ІС). Система обробки даних (СД) використовується в

системах аналізу даних. У цьому випадку застосовується технологія OLAP (Online Analytical Processing), яка дозволяє аналізувати дані у режимі реального часу (оперативний аналіз даних) [1].

Сховище даних повинно бути побудоване на основі бази даних, яка містить критеріальну і чисельну інформацію про відкриті сторінки на Facebook, Instagram та Twitter, що належать адміністраторам проекту Ecotwins.

Сховище даних містить інформацію про пост, тип посту, найменування соц мережі та час – це мають бути виміри сховища. Таблиця фактів має містити інформацію про час посту у відповідній соціальній мережі, його тип, кількість вподобань, коментарів та поширень.

Сховище даних надає всю необхідну інформацію для аналізу даних для відповідей на питання:

- в який час найкраще публікувати записи? (Час, день тижня)
- чи впливають наявність геолокації та відміток людей на загальну кількість перегляду запису?
- В якій соціальній мережі найкраща активність користувачів?
- Який контент найбільш сприйятливий для відвідувачів?
- Які дії найбільше роблять користувачі при відвідуванні вебсайту?

1.2 Аналіз предметної області

Аналіз предметної області, проведений для підготовки до подальшого проектування бази даних, має на меті створити єдиний погляд на інформацію, що обробляється в цій області. Це включає в себе розгляд структури даних, а також встановлення правил для зберігання та обробки цих даних, які потім відобразяться у визначенні функцій і завдань проектування.

Аналіз предметної області в рамках розробки інформаційних систем включає в себе виділення як основних, так і допоміжних бізнес-процесів [17], які потрібні для успішної реалізації системи аналізу. Водночас, цей процес

допомагає визначити бізнес-елементи та структури даних, які будуть використовуватися для обробки інформації.

Зазначені умови вимагають, щоб під час моделювання бази даних розробник брав до уваги не лише документи, що використовуються в певній галузі, а також оточення кожного бізнес-процесу і функцій. Це означає враховувати бізнес-елементи, об'єкти даних, осіб, які здійснюють обробку інформації, власників процесів та функцій, інші функції, які ініціюють або завершують події, та інші компоненти.

Детальний аналіз бізнес-процесів і функцій надає повний огляд процесів, що відбуваються в конкретній галузі і полегшує розуміння завдань, які необхідно виконати при створенні бази даних. Серед таких завдань входять моделювання структури бази даних, визначення правил посилальної цілісності, розробка процедур обробки і представлення даних, а також відповідь на запити користувачів [17].

Для найкращого висвітлення суті аналізу предметної області в дослідженні застосовуються моделі у вигляді графічних схем.

Діаграма варіантів використання (або Use Case Diagram) в рамках мови моделювання UML (Unified Modeling Language) є інструментом для представлення відносин між акторами і прецедентами, а також входить до структури моделі прецедентів. Вона служить для концептуального опису системи [17].

При роботі з варіантами використання було дотримано наступних правил:

- кожен прецедент відноситься як мінімум до одного актора;
- кожен прецедент має ініціатора;
- кожен прецедент призводить до відповідного результату.

Діаграма прецедентів для системи аналізу сторінок в соц.мережах та на вебсайті з позиції активності користувачів наведена на рис.1

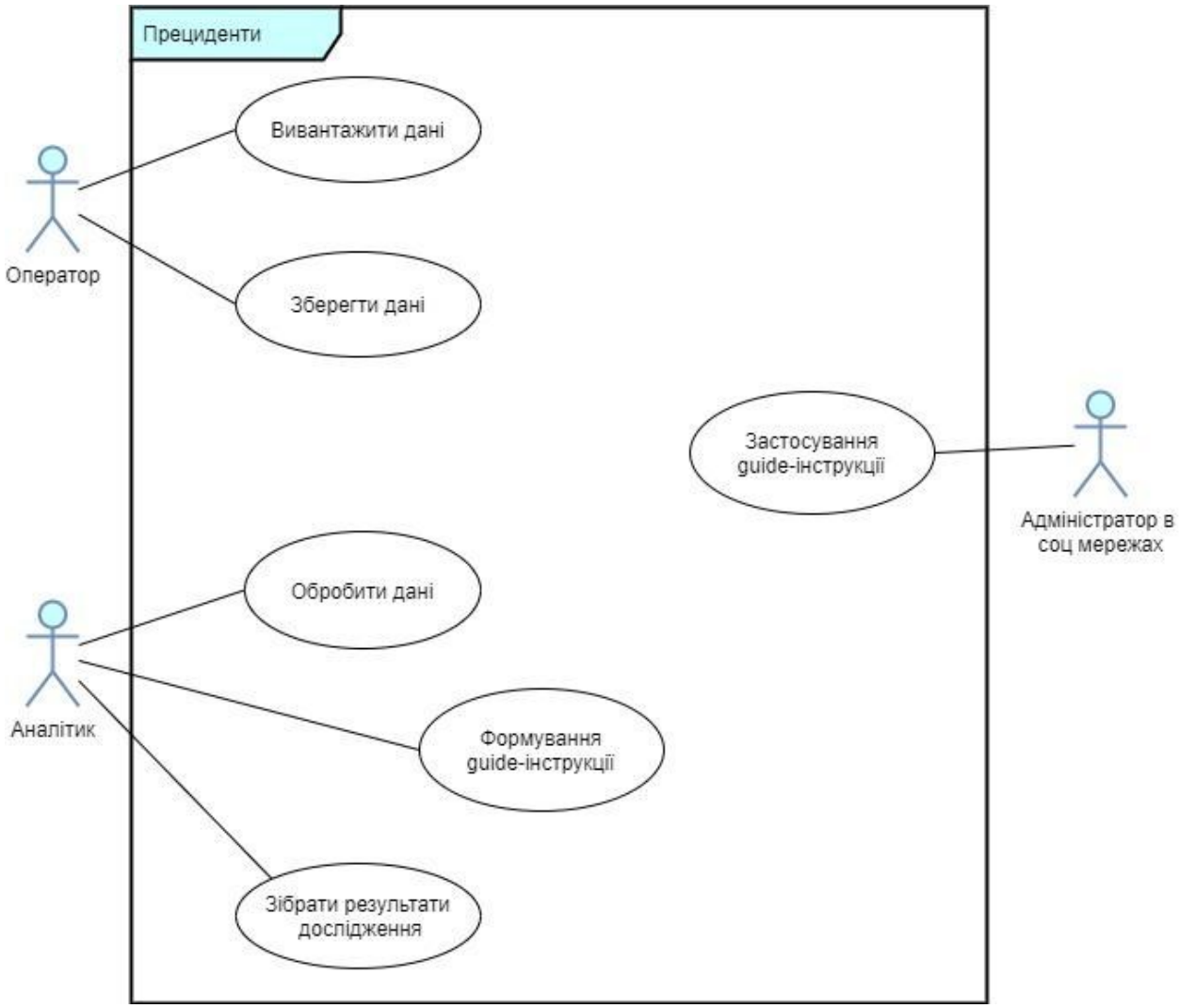


Рис. 1 Діаграма прецедентів

З діаграми видно, що з системою працюють три основні актори: оператор, аналітик та адміністратори сторінок в соціальних мережах: Facebook, Twitter(X) та Instagram та на вебсайті. Оператор працює із внесенням та збереженням даних в систему, задача аналітика опрацювати ці дані, створити аналітику та донести свої результати до користувача. В свою чергу, адміністратор сторінки має дотримуватись правил, що створив аналітик у інформаційній довідці (додаток Г).

Кожен з прецедентів, вказаних в діаграмі, має свої варіанти використання, їх детальний опис наведено в таблиці 1.

Таблиця 1

Опис прецедентів

Основний актор	Назва прецеденту	Опис прецеденту
Оператор	Вивантажити дані	Оператор займається завантаженням даних в систему.
Оператор	Зберегти дані	Оператор має зберігати завантажені дані.
Аналітик	Обробити дані	Аналітик займається обробкою та аналізом даних.
Аналітик	Акумулювати результати дослідження	Аналітик має зробити висновки щодо аналізу та візуалізувати результати дослідження.
Аналітик	Формування guide-інструкції	Результати дослідження мають бути викладені в інструкції, якою буде користуватися адміністратор спільноти.
Адміністратори спільнот в соц. мережах	Застосування guide-інструкції	Адміністратори спільнот користується порадами аналітика задля покращення статистики спільноти в соціальних мережах Facebook, Twitter та Instagram.

1.3 Проектування системи

Процес проектування системи для аналізу активності користувачів у соціальних мережах і на веб-сайті розпочався з чіткого визначення мети цієї системи. Головним завданням для досягнення успішності проекту є забезпечення надійного функціонування системи як на момент її запуску, так і протягом всього періоду експлуатації. Це охоплює такі аспекти:

- Необхідна функціональність системи та її здатність адаптуватися до постійно змінюючихся умов;
- Пропускна здатність системи, яка визначає, скільки даних вона може обробляти;

- Час реакції системи на запити користувачів;
- Надійність системи, її здатність працювати без відмов та бути доступною для обробки користувальницьких запитів;
- Простоту в експлуатації та підтримці системи;
- Забезпечення відповідного рівня безпеки в системі [26].

Ефективність системи визначається її продуктивністю, яка є одним із ключових факторів. Вдале проектування виступає основою для створення високопродуктивної системи.

Під час проектування системи аналізу активності користувачів у соціальних мережах Facebook,Twitter(X), Instagram та на вебсайті були враховані три основні аспекти:

- Проектування структури даних, які будуть використовані в базі даних.
- Проектування програм, екранних форм та звітів для обробки та відображення даних.
- Аналіз середовища та технології, включаючи мережеву топологію, конфігурацію апаратних ресурсів, вибір архітектури (файл-сервер чи клієнт-сервер), паралельну обробку, розподілену обробку даних тощо.

Отже, наступним етапом була розробка моделі та структури системи. Архітектура системи визначається як сукупність взаємозв'язків та взаємодій між її основними функціональними компонентами та засобами, які забезпечують їх взаємодію з користувачем та персоналом системи [5, 26]. Важливою характеристикою системної архітектури є спроектовані зв'язки та взаємодія між всіма компонентами системи.Архітектуру системи зображено на рис.2.

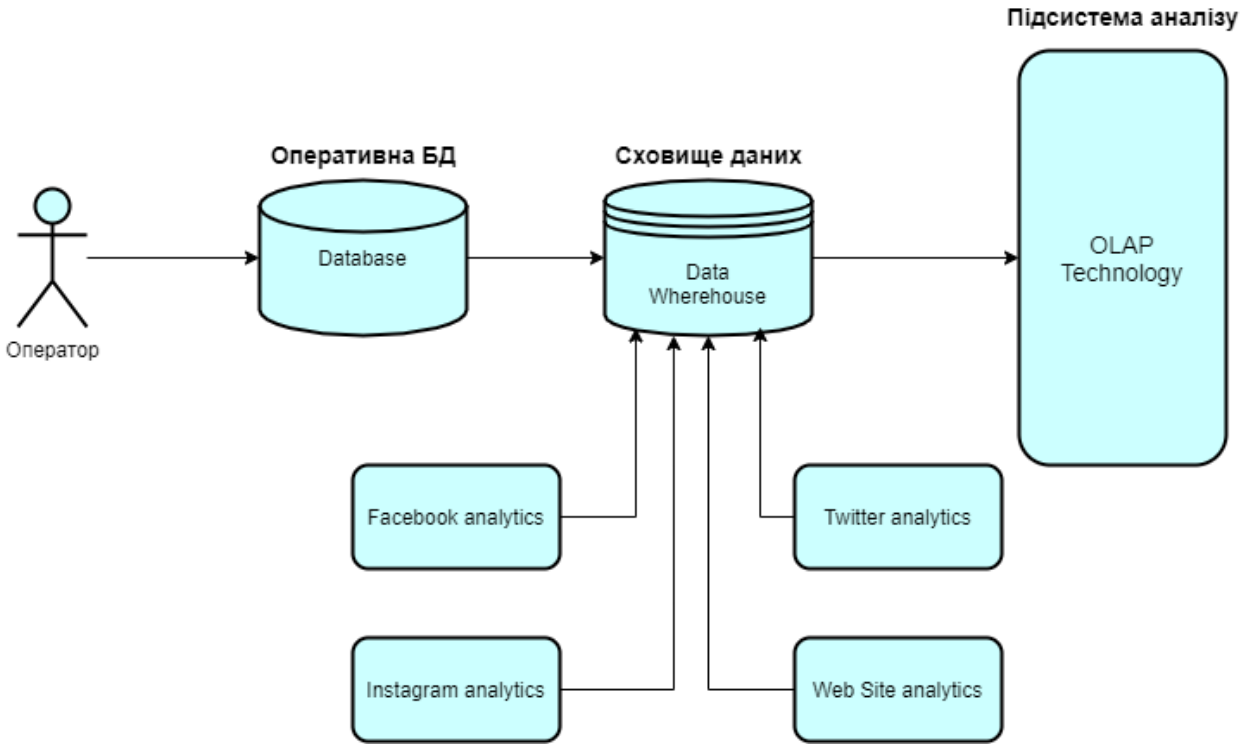


Рис. 2 Топологія системи

Інформаційно-аналітична система аналізу сторінок в соціальних мережах з позиції активності користувачів складається з оператора, який вносить дані в систему та відправляє запити; бази даних, яка містить всю потрібну інформацію про сторінки в соц.мережах та про вебсайт; сховище даних, в якому містяться дані з бази та додаткового джерела Facebook analytics (дані, що завантажуються до системи за допомогою Facebook Business Suite); сховище даних, в якому містяться дані з бази та додаткового джерела Google Analytics; сховище даних, в якому містяться дані з бази та додаткового джерела witter API [6]. В сховищі даних не обов’язково міститься вся інформація з баз даних, це можуть бути лише обрані таблиці або деякі значення і т.д... В результаті ми маємо підсистему аналізу - OLAP технологію, за допомогою якої буде відбуватися аналіз даних, що зберігається в сховищі, та візуалізація результатів роботи.

2 МЕТОДИ ТА ТЕХНОЛОГІЇ АНАЛІЗУ

2.1 Аналіз методів обробки даних

Обробка даних включає в себе різноманітні методи, які використовуються для структурування та аналізу існуючої інформації. Задачі аналізу даних різноманіті, але у рамках цього дослідження були розглянуті методи, які ефективно впораються з завданнями структурування даних з великою кількістю різноманітних параметрів [7].

Наприклад, для привернення більшої уваги цільової аудиторії, розумно провести сегментацію реакцій читачів сторінок за різними параметрами, такими як кількість вподобань, поширень, коментарів та переглядів. Для досягнення цієї мети можна використовувати різні математичні методи, які допоможуть виявити закономірності в даних. Наприклад, при аналізі кількісних реакцій на публікації можна виявити, що публікації зі схожим типом контенту мають схожі реакції від аудиторії.

Дуже часто формальні методи аналізу даних можуть внести несподівані нові знання. Наприклад, під час вивчення обсягів тексту публікацій може стати зрозумілим, що більшість читачів реагують на дуже короткі тексти (до 80 символів) або на дуже об'ємні (1100-1300 символів). Таким чином, перед вибором методу обробки даних важливо ретельно розглянути, які дані будуть використовуватися для аналізу та яка кількість даних є необхідною.

Для аналізу використовуються дані чотирьох основних типів:

- Кількісні дані, які включають в себе числа, такі як кількість лайків, коментарів, репостів і переглядів.
- Інтервальні дані, які охоплюють об'єм тексту, що використовується в публікаціях.
- Рангові дані, які описують порядок або рівень важливості, такі як рейтинги геолокацій, користувачів та налаштування цільової реклами.

- Номінальні дані, що визначають різні категорії або типи мультимедійних даних, доданих до публікацій.

Всі дані, які входять в один із цих типів, можуть бути аналізовані за допомогою формальних методів. Набір даних може бути відповідно представлений як комбінація наведених типів.

Також важливо відзначити, що обсяг даних, які обробляються системою, відносно невеликий. Так, сторінки та вебсайт проекту Ecotwins існують лише протягом останніх 2 років і мають від 48 до 200 публікацій.

2.2 Загальні поняття з напрямку OLAP-технології

OLAP, або Online Analytical Processing, є методом обробки даних, який використовується для аналізу великих обсягів інформації, організованих у багатовимірній структурі. Реалізація цієї технології зазвичай включає в себе використання компонентів з області бізнес-інтелігенції. [1].

Термін "OLAP" був винайдений Едгаром Коддом, і в 1993 році він сформулював "12 правил аналітичної обробки в реальному часі," аналогічно до раніше сформульованих "12 правил для реляційних баз даних."

Однією з головних мотивацій для використання OLAP для обробки запитів є швидкість. Реляційні бази даних зазвичай зберігають сутності в окремих таблицях, які нормалізовані. Ця структура досить зручна для операційних баз даних (систем OLTP), але складні запити до багатьох таблиць виконуються в них відносно повільно [17].

OLAP-структура, утворена на основі робочих даних, отримала назву "OLAP-куб". Цей куб формується шляхом об'єднання таблиць за допомогою схеми "зірка" чи схеми "сніжинка" (іноді відомої як "крижинка"). У центрі будь-якої такої схеми розташована таблиця фактів, яка містить ключові дані, які використовуються для створення запитів. Різні таблиці з вимірами приєднуються до цієї таблиці. Ці вимірювальні таблиці вказують, як дані можуть бути агреговані і аналізовані. Кількість можливих агрегацій

визначається кількістю способів, якими початкові дані можуть бути ієрархічно представлені.

OLAP-куб включає в себе основні дані та інформацію про агрегати. Цей куб теоретично містить всю необхідну інформацію для відповіді на будь-які запити. Проте, при великій кількості агрегатів, розрахунки зазвичай виконуються тільки для певних вимірювань, інші розраховуються "на вимогу".

Існують три основних типи OLAP:

- Багатовимірна OLAP (Multidimensional OLAP - MOLAP);
- Реляційна OLAP (Relational OLAP - ROLAP);
- Гібридна OLAP (Hybrid OLAP - HOLAP).

MOLAP - це класична форма OLAP, яку часто просто називають OLAP. Вона базується на створенні сумаризованої бази даних і створює потрібну багатовимірну схему даних зі збереженням як основних даних, так і агрегатів.

ROLAP працює безпосередньо з реляційною базою даних. Факти та таблиці з вимірами зберігаються у реляційних таблицях, і для збереження агрегатів створюються додаткові реляційні таблиці.

HOLAP використовує реляційні таблиці для зберігання основних даних і багатовимірні таблиці для агрегатів.

Особливим випадком ROLAP є "ROLAP реального часу" (Real-time ROLAP - R-ROLAP). У відміну від ROLAP в R-ROLAP не створюються додаткові реляційні таблиці для зберігання агрегатів. Агрегати обчислюються в момент запиту. При цьому багатовимірний запит до OLAP-системи автоматично перетворюється в SQL-запит до реляційних даних.

Кожен метод зберігання даних має свої переваги, і їх оцінка різниться в залежності від виробника. MOLAP найбільше підходить для невеликих обсягів даних, оскільки він швидко обчислює агрегати та надає оперативні відповіді, але при цьому може призводити до значних обсягів даних. ROLAP вважається більш масштабованим рішенням та є більш економічним з точки зору обсягу зберігання, але має обмеження в аналітичній обробці. HOLAP розташовується

між цими двома підходами, добре масштабується та допомагає подолати деякі обмеження. Архітектура R-ROLAP дозволяє проводити багатовимірний аналіз OLTP-даних в режимі реального часу [2].

Складність використання OLAP полягає в необхідності генерувати запити, вибирати вихідні дані та розробляти схеми. Тому багато продуктів OLAP постачаються з великою кількістю заздалегідь налаштованих запитів. Ще однією проблемою є базові дані, які повинні бути консистентними та без суперечок.

Термін "OLAP" використовується для опису моделі представлення даних та відповідної технології їх обробки в сховищах даних. У фреймворку OLAP використовується багатовимірне представлення агрегованих даних з метою надання швидкого доступу до стратегічно важливої інформації для поглибленого аналізу. Додатки OLAP повинні відповідати таким ключовим властивостям:

- Використання багатовимірного подання даних.
- Підтримка складних розрахунків.
- Правильне врахування часового фактору.

Системи OLAP базуються на сховищах даних і отримують актуальну інформацію від них, що сприяє забезпеченню цілісності корпоративних даних [2].

Отже, OLAP - це технологія для оперативного аналізу даних, яка використовує методи та інструменти для збору, зберігання і аналізу багатовимірних даних з метою підтримки процесів прийняття рішень. Основне призначення OLAP-систем полягає в підтримці аналітичних операцій та у можливості користувачів-аналітиків формулювати різноманітні запити. Головна ціль OLAP-аналізу полягає в перевірці гіпотез [1].

На рис.3 показана схема діяльності інформаційних систем, основаних на OLAP технології.

Так, процес функціонування системи розпочинається з введення інформаційних даних через підсистему введення (OLTP), де оператор системи вводить цю інформацію. Далі інформація надсилається до підсистеми збереження даних, де вона потрапляє до сховища даних. Тут інформація сортується та набуває структури, готової для подальшого аналізу. Важливо відзначити, що система може отримувати інформацію із сторонніх джерел, відомих як зовнішні джерела даних. Інформація, зібрана в підсистемі збереження даних, переходить до підсистеми аналізу, яку керує аналітик. На цьому етапі можливий пошук нових знань інформації, що вже наявна, та формування звітів і результатів аналізу.

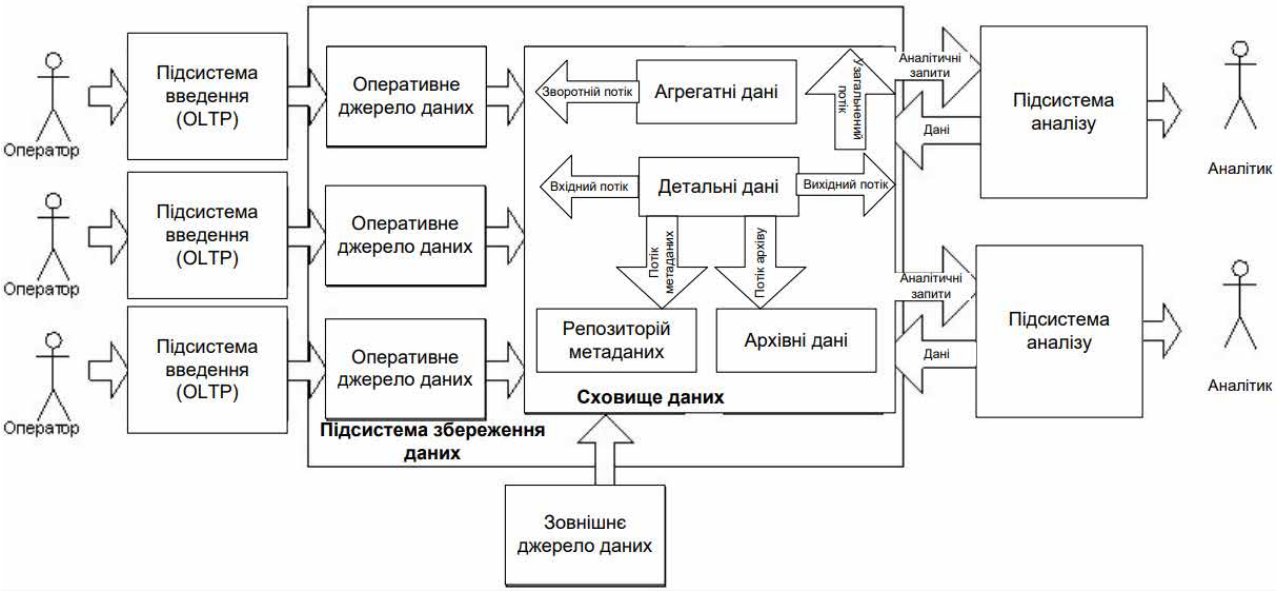


Рис. 3 Категорії даних та інформаційні потоки в OLAP [2].

В OLAP-системах дані поділяються на три основні категорії: детальні дані, агреговані дані та метадані.

Детальні дані: Це дані, які переносяться безпосередньо з підсистеми OLTP (Online Transaction Processing). Вони відповідають окремим елементарним подіям і фіксуються в OLTP-системах. Детальні дані поділяються на:

Виміри: Це набори даних, які використовуються для опису окремих подій або об'єктів (наприклад, товар, продавець, покупець, магазин і т. д.).

Факти: Це дані, що відображають конкретні значення чи характеристики подій (наприклад, кількість проданих товарів, сума продажів і т. д.).

Агреговані (узагальнені) дані: Це дані, які створюються на основі детальних даних шляхом їх підсумовування за певними вимірами. Тобто це результати аналізу та узагальнення детальних даних.

Метадані: Це дані про дані, які містяться в системі обробки даних (СД).

Метадані можуть включати інформацію про:

- Об'єкти предметної області, дані про які містяться в СД.
- Категорії користувачів, які використовують дані в СД.
- Місця та способи зберігання даних в СД.
- Дії, які виконуються над даними в рамках аналітичного процесу.
- Час виконання різноманітних дій над даними.
- Причини виконання різних дій над даними [1, ст.26].

Для забезпечення коректної роботи OLAP-систем потрібно виконати кілька етапів, які допоможуть зібрати інформацію в системі обробки даних (СД). Ця інформація буде використовуватися для аналізу процесів у корпорації. Основні етапи включають в себе [5]:

- Вилучення та перетворення даних.
- Очищення даних.
- Завантаження даних до СД.
- Оновлення даних в СД.
- Управління метаданими.

2.3 Моделювання сховища даних

Сховище даних, також відоме як data warehouse, є спеціально створеним та інтегрованим набором даних, орієнтованим на певну предметну область.

Воно має можливість зберігати дані з урахуванням їхньої хронології та надає комплексний доступ до надійної інформації для оперативного аналізу та прийняття рішень. Основним принципом концепції сховища даних є розрізнення між інформацією, яку використовують в системах оперативної обробки даних (OLTP), та тією, яку використовують в системах підтримки прийняття рішень (СППР) [3, 8].

Переміщення даних з OLTP-системи в сховище даних виконується таким чином, щоб під час створення звітів та проведення аналізу OLAP не використовувалися ресурси транзакційної системи і не завдали шкоди її стабільності. Існують два способи оновлення даних в сховищі:

1. **Повне оновлення даних в сховищі:** спочатку старі дані видаляються, а потім нові дані завантажуються. Цей процес відбувається з певною періодичністю, і це може призвести до того, що актуальність даних в сховищі трохи відставатиме від OLTP-системи.
2. **Інкрементальне оновлення:** оновлюються лише ті дані, які зазнали змін в OLTP-системі.Існують два основних архітектурних напрямки - нормалізовані сховища даних і сховища з вимірами.

Дані у нормалізованих сховищах розміщені в таблицях, які відповідають третій нормальній формі і є спрощеними у процесі створення та управління.

Недоліки нормалізованих сховищ полягають у великій кількості таблиць, яка виникає через процес нормалізації. Це означає, що для отримання інформації користувачам часто доводиться взаємодіяти з багатьма таблицями одночасно, що може призвести до погіршення продуктивності системи. Для подолання цього недоліку використовують денормалізовані таблиці, які називаються "вітринами даних." Вони використовуються для створення звітів і аналітичних форм. При роботі з великими обсягами даних можуть використовувати кілька рівнів таких "вітрин" або "сховищ" для полегшення і підвищення продуктивності обробки і аналізу інформації.

Після вивчення наукової літератури та попереднього аналізу зібраних даних, було визначено, що найкращим рішенням буде створення нормалізованого сховища даних зі структурною схемою "зірка".

3 РОЗРОБКА СИСТЕМИ АНАЛІЗУ

3.1 Опис вузлів системи, які поставляють дані по сховищу

Зібрані дані потрапляють до бази даних та розподіляються за сімома таблицями. На рис.4 представлено ER діаграму бази даних, створену в додатку ER-win.

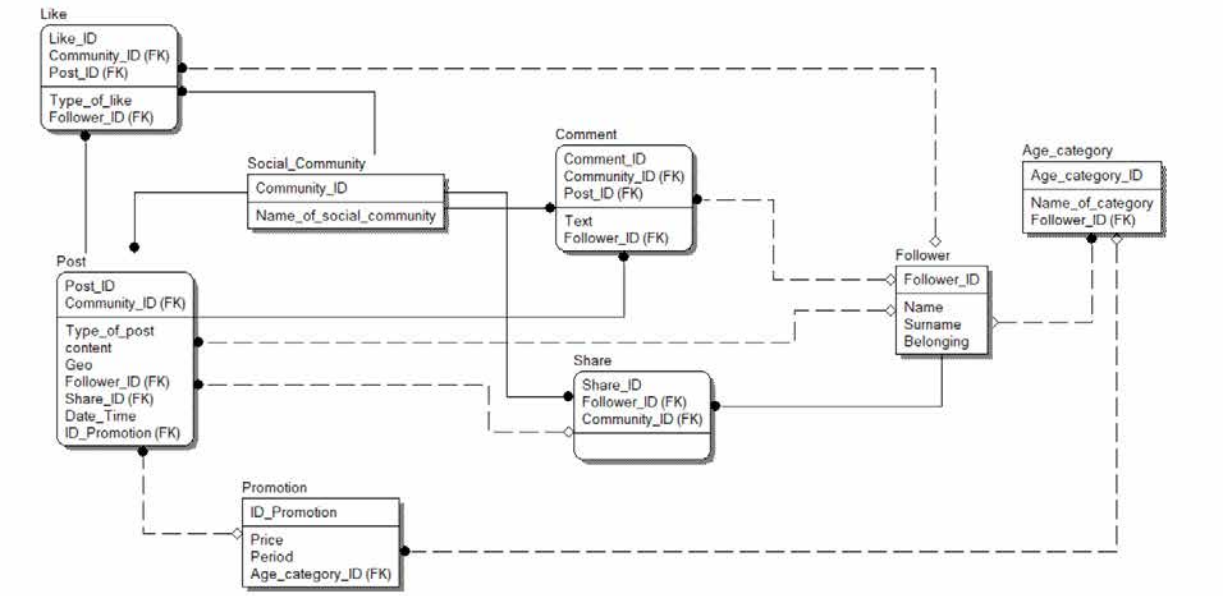


Рис. 4 ER діаграма бази даних

Розроблена база даних зберігає в собі будь-яку подію що відбувається на сторінках спільнот. Так, поставивши вподобання на записі, система зберігає інформацію коли відбулась подія (рік, місяць, дата, день тижня, час), на якому записі, в якій соцмережі та яким користувачем.

Варто зауважити, що Facebook зберігає деталізовану інформацію з приводу часу публікації один місяць, Twiiter(X) зберігає деталізовану інформацію постійно, Instagram зберігає деталізовану інформацію протягом 7 днів. Тому, якщо адмінам необхідна буде така інформація, вивантаження даних має бути: для Фейсбуку-щомісяця або частіше, для Твітеру(X) будь-коли, однак рекомендацією є вивантаження інформації щомісяця, а для інстаграму варто моніторити і вивантажувати дані кожних 7 днів,при необхідності.

Сім даних таблиць і фіксований набір атрибутів повністю відповідають вимогам системи і є достатніми для проведення ретельного аналізу. Таблиця "Time_attribute" містить всю необхідну інформацію про час подій, "Like" показує, в якій спільноті, на якому записі, ким і о котрій годині було залишено вподобання певного типу. Таблиці "Post" та "Share" створені на аналогічних принципах. Таблиця "Community" містить інформацію про назву соціальної мережі, і таблиця "Follower" вказує на тип користувача (адміністратор, відвідувач або адміністратор спільноти).

Аналіз даних оперативних систем важко проводити без використання сховищ даних, і це пояснюється декількома причинами. По-перше, дані можуть бути розрізнені та зберігатися у різних форматах у різних системах керування базами даних (СКБД) та в різних частинах корпоративної мережі. Це стосується як великих корпорацій, так і навіть невеликих підприємств, де може бути кілька джерел даних.

У зв'язку з цим, для проведення дослідження було розроблено та впроваджено сховище даних, в якому інформація зберігається у єдиному форматі, не є розрізненою, структурованою в одному місці та готовою до подальшого використання аналітиком. Структура розробленого та реалізованого сховища даних представлена на рис.5.

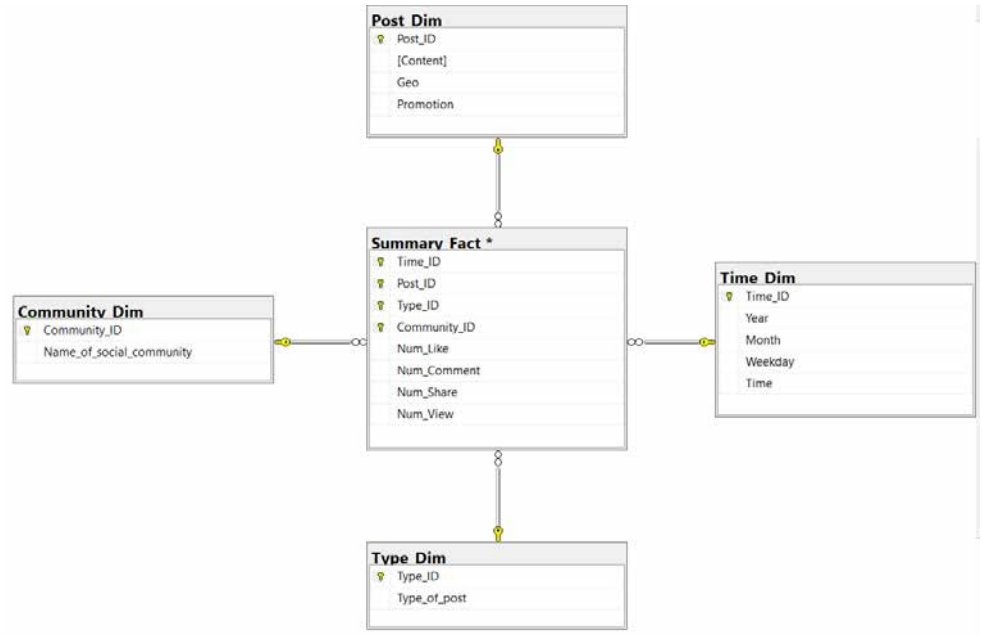


Рис. 3 Структура сховища даних

Для зберігання даних було використано структурну схему «зірка», де розміщено 4 виміри та одна таблицю фактів:

- Post_Dim – містить інформацію про наповнення посту (кількість символів тексту), чи має пост прив’язку до георозташування, чи має пост рекламну проплату для просування;
- Community_Dim – вимір містить назву спільноти в соц. мережі;
- Type_Dim – вимір містить інформацію, яка медіа використовується в даному пості;
- Time_Dim – фіксує час подій;
- Summary_Fact – до таблиці фактів мігрують всі ключі з таблиць вимірів, крім цього ми зберігаємо кількість вподобань, коментарів, поширень та переглядів посту.

На цьому етапі пояснювальної записки необхідно детально показати кроки заповнення сховища даних. Для переміщення інформації з бази даних до сховища створено запити, що дозволяли інформації мігрувати (додаток Ґ).

Таким чином, сховище даних забезпечує аналітикам інформаційної системи інформаційні ресурси, які знаходяться в одному місці і мають чітку структуру. Це дозволяє нам переходити до наступного етапу дослідження, який включає в себе підготовку даних зі сховища для подальшого аналізу та обробки.

3.2 Механізм вилучення, обробки і передачі даних

3.2.1 Опис BI та створення в його середовищі проекту служби SSAS (побудова розгорнутого куба). Business Intelligence (BI), або бізнес-інтелігенція, представляє собою набір комп'ютерних методів та інструментів для організацій. Вона спрямована на перетворення ділової інформації, зібраної під час транзакцій, в зрозумілу форму для аналізу та надає засоби для роботи з такою обробленою інформацією великого обсягу.

Головною метою BI є інтерпретація великої кількості даних, зосереджуючись на ключових факторах ефективності та моделюючи результати різних сценаріїв, щоб відстежувати наслідки прийнятих рішень.

Business Intelligence (BI) включає в себе різноманітні бізнес-рішення, охоплюючи спектр від операційних до стратегічних. Операційні рішення часто включають в себе аспекти, такі як позиціонування продукту і ціноутворення. Стратегічні бізнес-рішення, натомість, охоплюють питання пріоритетів, цілей та загальних напрямків розвитку компанії.

Ефективність BI найкраще виявляється тоді, коли він поєднує дані з зовнішніх джерел, таких як дані ринку, на якому діє компанія, з внутрішніми даними компанії, такими як фінансова та виробнича інформація. Ця комбінація зовнішніх і внутрішніх даних надає компанії більш повний інсайт у її діяльність, а також забезпечує структуровані дані для аналізу, які неможливо отримати, опираючись лише на одне джерело інформації [7].

Інші вчені розглядають термін "Business Intelligence" (BI) як метафору, яка не має прямого перекладу або однозначного визначення. Вони визначають цей термін як ієрархічний та синергетичний комплекс концепцій, технологій та програмних інструментів для аналізу початкових даних та візуалізації результатів цього аналізу з метою підтримки процесу ухвалення рішень.

Business Intelligence поєднує в собі технології, які працюють з реляційними та нереляційними (NoSQL) базами даних, а також використовують передові технології штучного інтелекту та традиційної статистики разом із інструментами візуалізації результатів аналізу.

Системи Business Intelligence (BI) зазвичай використовують дані, які зберігаються у сховищі даних. Вони включають наступні три основні категорії функцій:

- Можливість інтеграції.
- Представлення інформації.
- Аналіз даних.

Для розробки системи аналізу використовувалася **SQL Server Business Intelligence.** У цьому середовищі створювався куб та налаштовувалися потоки даних для інтеграції інформації з різних джерел у цей куб. Для побудови кубу був створений проект типу Analysis Services (рис.6).

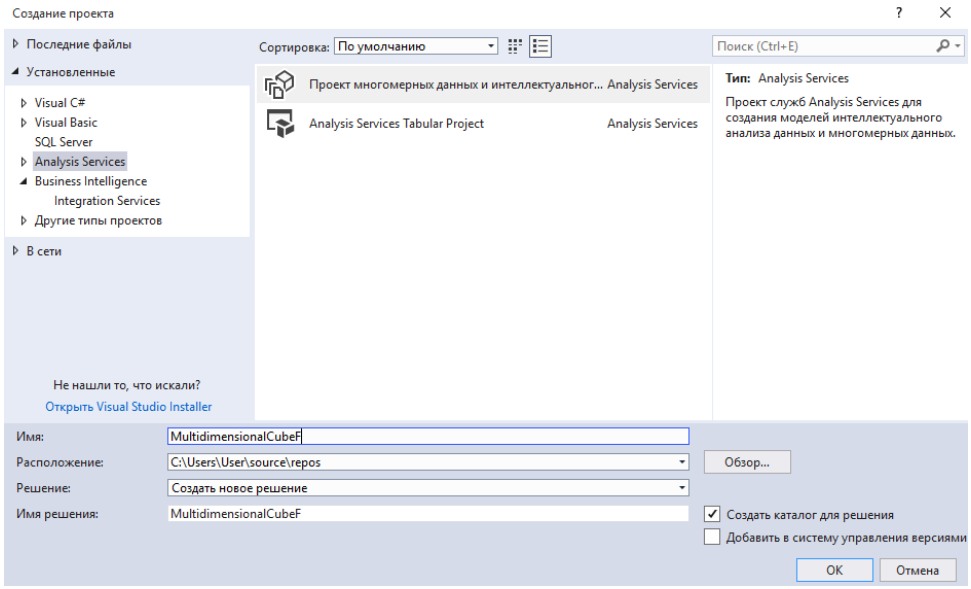


Рис. 6 Створення проекту для розгортання кубу

Наступним етапом є підключення до джерела даних Data Source та створення Data Source View (рис. 7). Під Data Source View розуміється витяг з джерела, який буде використовуватися для наповнення сховища. В Data Source View можуть входити як таблиці, так і представлення (views) з реляційної бази даних - джерела інформації.

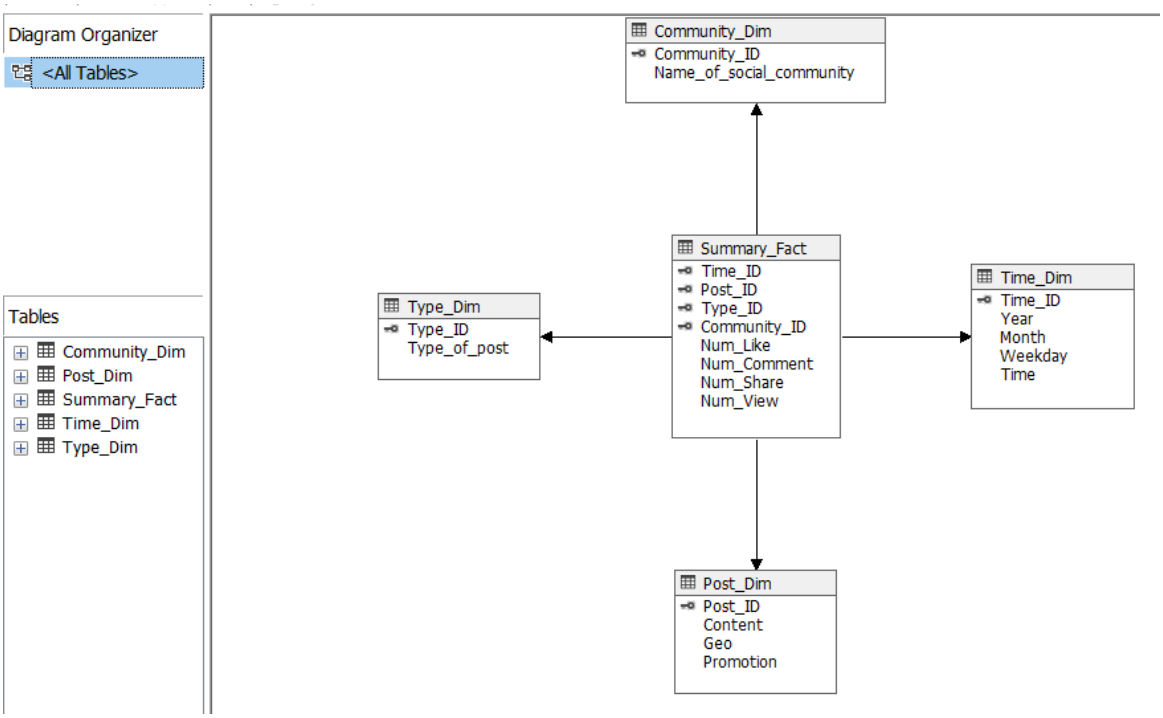


Рис. 7 Результат Data Source View

Далі, наступним кроком є розгортання кубу. На рис. 8 представлені всі розгорнуті виміри, а на рисунку 9 показана повномасштабна схема кубу. Більш докладні інструкції щодо створення розгорнутого кубу наведені в додатку Д.

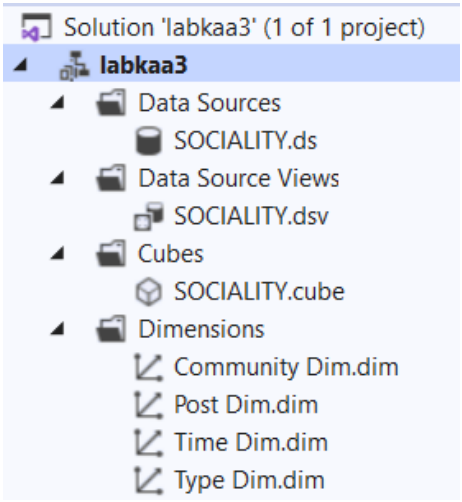


Рис. 8 Результат розгортання кубу

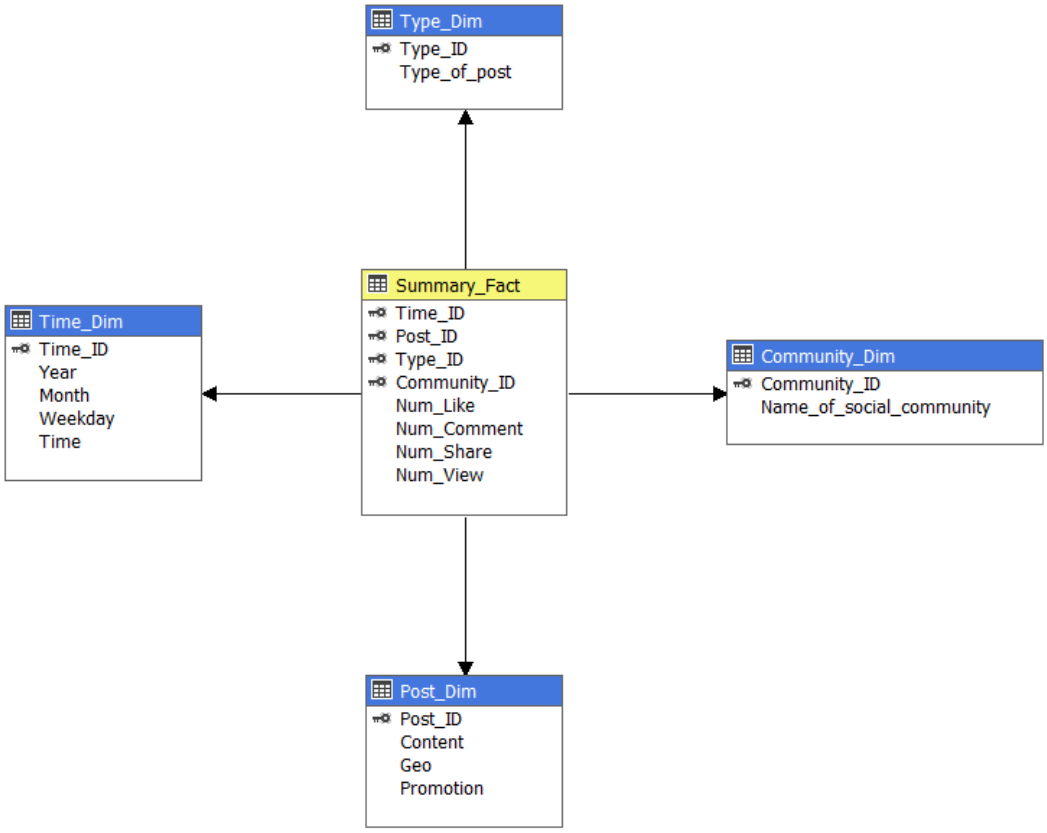


Рис. 9 Розгорнутий куб

Створення структури розгорнутого гіперкубу (також відомого як "куб даних") дійсно є потужним інструментом для зберігання та аналізу великих обсягів даних. Ця структура дозволяє організувати інформацію в багатовимірний куб, де кожен вимір представляє різні аспекти даних, які підлягають аналізу.

Основні переваги створення структури гіперкубу включають:

1. Швидкий доступ до даних: Гіперкуб дозволяє швидкий доступ до даних за допомогою різних комбінацій вимірів. Це полегшує вибірку та агрегацію даних, що значно прискорює аналіз.
2. Можливість проведення зрізів даних: Ви можете створювати зрізи даних для досліджень та аналізу конкретних аспектів інформації.
3. Здатність виявлення тенденцій і закономірностей: Аналіз гіперкубу допомагає виявити зв'язки між різними вимірами даних і розкрити тенденції, які можуть бути непомітними в іншій формі організації даних.

4. Можливість багаторазового використання: Розгорнутий гіперкуб можна використовувати для різних завдань і досліджень, оскільки він має гнучку структуру.
5. Підтримка прийняття рішень: Гіперкуби дозволяють аналізувати дані і вирішувати різноманітні завдання, такі як прогнозування та прийняття рішень на основі доступної інформації.

Створення структури гіперкубу може вимагати великої роботи з обробки та інтеграції даних, а також визначення вимірів і вимог для дослідження. Проте, якщо вона правильно розроблена, ця структура може бути потужним інструментом для аналітики та вирішення комплексних завдань у сфері обробки даних.

3.2.2 Реалізація отриманих даних за допомогою Data Flow. На цьому етапі дослідження будемо заповнювати гіперкуб даними, які в подальшому будемо аналізувати. Отримання даних з джерела та заповнення згенерованого кубу було виконане за допомогою Data Flow. Для цього було створено проект служб SSIS (рис.10).

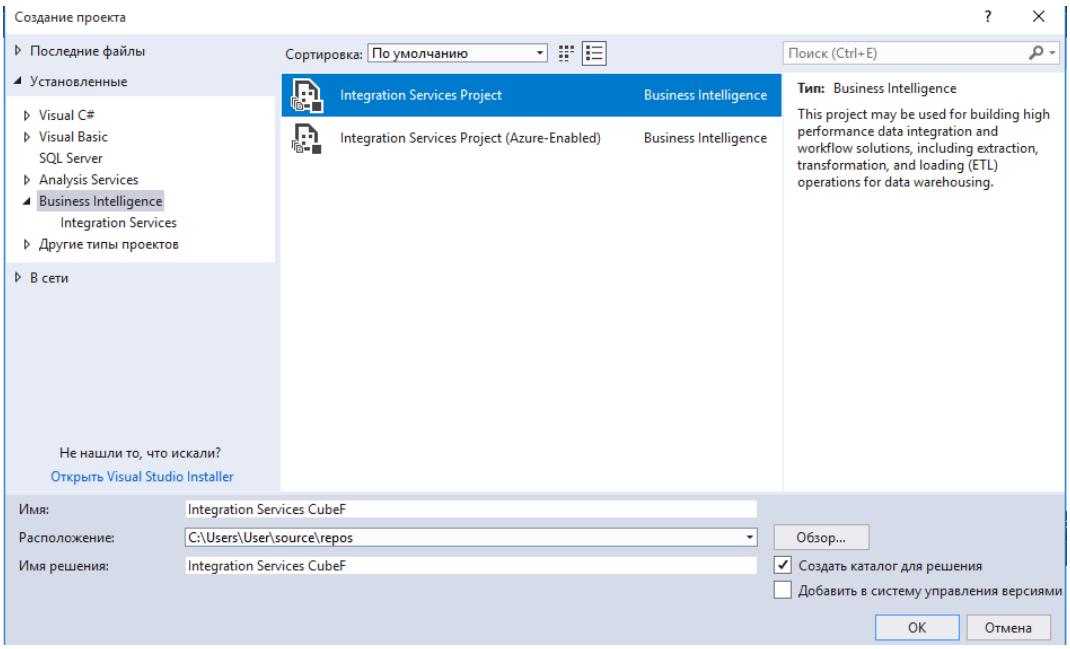


Рис. 10 Створення проекту для заповнення кубу даними

За допомогою служби SSIS, на основі процесів Data Flow, заповнено даними побудований куб. На рисунках 11 – 14 відображено результат

заповнення розгорнутого куба та відповідність стовпців джерела та приймача даних.

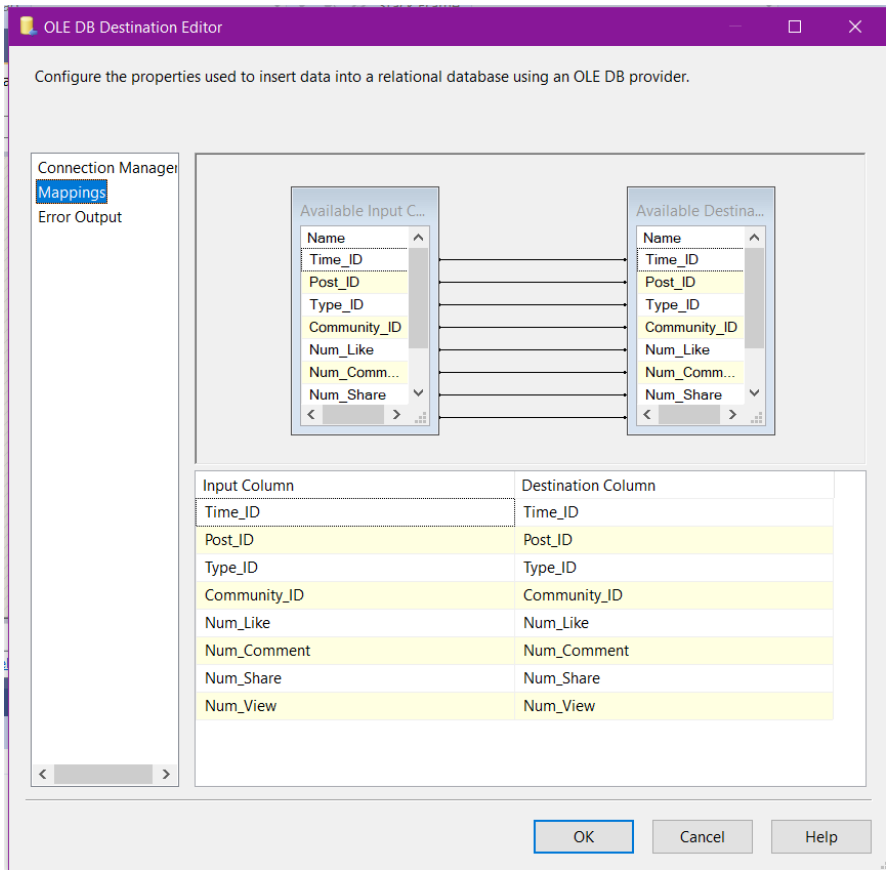


Рис. 11 Відповідність колонок

	Time_ID	Post_ID	Type_ID	Community...	Num_Like	Num_Com...	Num_Share	Num_View
▶	time1	Post01	Type4	Soc1	3	0	6	70
	time11	Post12	Type7	Soc1	0	1	0	57
	time12	Post10	Type4	Soc3	5	0	1	102
	time2	Post05	Type5	Soc3	6	1	2	93
	time2	Post08	Type4	Soc1	3	6	0	81
	time3	Post06	Type4	Soc1	0	0	5	65
	time4	Post09	Type4	Soc2	0	4	0	72
	time5	Post12	Type3	Soc2	0	0	2	57
	time6	Post04	Type4	Soc1	3	3	0	92
	time7	Post04	Type5	Soc3	5	0	3	92
	time8	Post01	Type4	Soc3	0	0	0	44
	time8	Post07	Type4	Soc2	3	0	0	79
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Рис. 12 Внесені дані

Потім створити задачу потоку даних для вимірів першої черги (рис.13), а в ній створити потоки даних для кожного виміру (рис.14) та факту (рис.15).

На скріншотах представлено запущений проект, де видно позитивний результат виконання.

Рис. 13 Потоки управління даних для наповнення таблиці-фактів.

Перший Потік:

Формування ключів факторів, яке складається з формування вибірки, та збереження даних в таблицю.

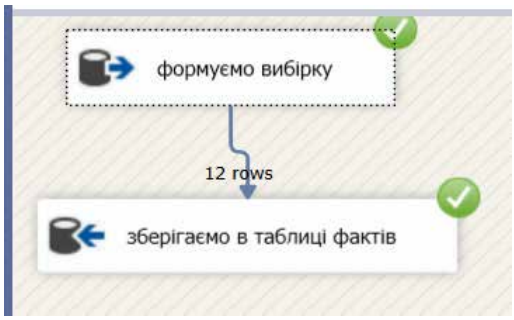


Рис. 14 Формування ключів фактів

Другий Потік:

Формування фактів, яке складається з формування запиту, обробки даних з запиту та збереження оброблених фактів в запиті.

Рис. 15 Формування фактів

3.3 Реалізація процедури аналізу даних в розробленій системі

3.3.1 Побудова звітності в середовищі BI. Звіт Power BI – це інтерактивний засіб для візуалізації та аналізу даних. Power BI - це платформа від Microsoft для бізнес-аналізу, яка дозволяє вам об'єднувати дані з різних джерел, створювати візуалізації та розповсюджувати звіти для прийняття бізнес-рішень.

Основні характеристики звіту Power BI включають:

1. Візуалізація даних: Звіти Power BI надають можливість створювати різні типи візуалізацій, такі як графіки, діаграми, таблиці, карти тощо. Ці візуалізації допомагають краще розуміти дані та їх зв'язки.
2. Інтерактивність: Ви можете додати інтерактивність до вашого звіту, дозволяючи користувачам взаємодіяти з даними. Наприклад, вони можуть фільтрувати дані, вибирати конкретні елементи для аналізу або динамічно змінювати періоди часу.
3. Багаторівневі сторінки: Звіти Power BI можуть включати кілька сторінок, кожна з яких може містити різні візуалізації для відображення різних аспектів даних. Це дозволяє краще організувати і відображати інформацію.
4. Автоматичне оновлення: Ви можете налаштувати автоматичне оновлення даних у звіті, щоб завжди мати актуальну інформацію.
5. Спільний доступ: Звіти Power BI можна розповсюджувати і спільно використовувати з іншими користувачами, що дозволяє команді або організації працювати з одними та самими даними та аналізувати їх разом.
6. Спеціалізовані інструменти: Power BI має ряд інструментів для аналізу даних, таких як DAX (Data Analysis Expressions), які допомагають виконувати розрахунки та обчислення на основі даних.

Звіти в Power BI створюються на основі конкретного набору даних. Кожна візуалізація у звіті відображає частину цих даних. Працюючи з цими

даними, ви можете додавати, видаляти їх, змінювати типи візуалізацій і застосовувати фільтри та зрізи для глибокого аналізу даних та відповіді на питання. Ці звіти, подібно до інформаційних панелей, надають широкі можливості налаштування та взаємодії з даними, і вони автоматично оновлюються при зміні вихідних даних.

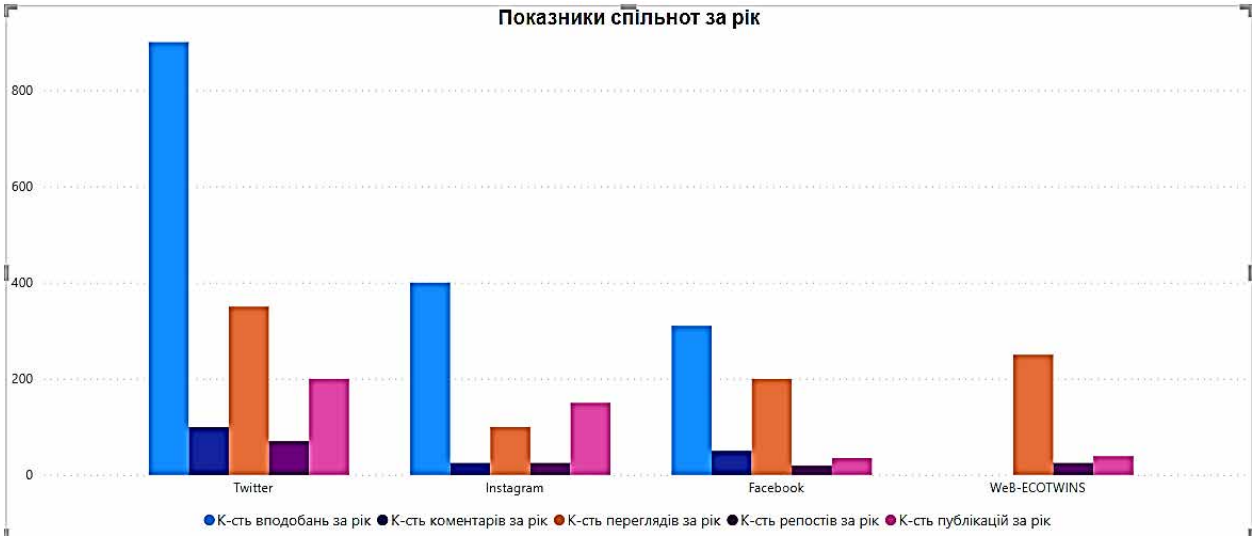


Рис. 16 Показники спільнот

Завдяки формуванню звітів в Power BI, знаходимо відповіді на питання, що задали при постановці завдання. На рис. 16 бачимо числову складову основних показників спільнот в соцмережах та на сайті за останній рік. В результаті можна подивитись реакцію відвідувачів на певну кількість записів у тій чи іншій соціальній мережі. Уточнюю, що дані було зібрано з жовтня 2022 по жовтень 2023 року.

На діаграмі бачимо, що найбільшу кількість вподобань, коментарів та переглядів було у сторінки проекту в соцмережі Twitter(X).

На іншій діаграмі (рис. 17) можна подивитись, в який день тижня було опубліковано найбільшу кількість постів в кожній соціальній мережі.

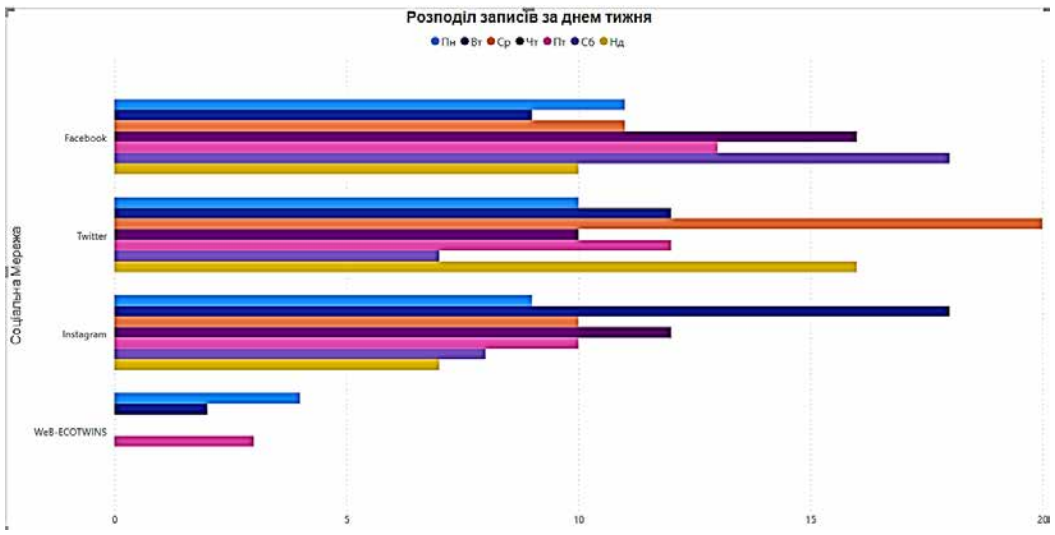


Рис. 17 Кількість записів за днем тижня

Для отримання більшої результативності у зборі даних, потрібно визначити коли(в який день) у сторінок найкраща реакція від відвідувачів, тому на наступному зображенні показуємо записи, де кількість вподобань, коментарів та поширень найбільша (рис. 18).

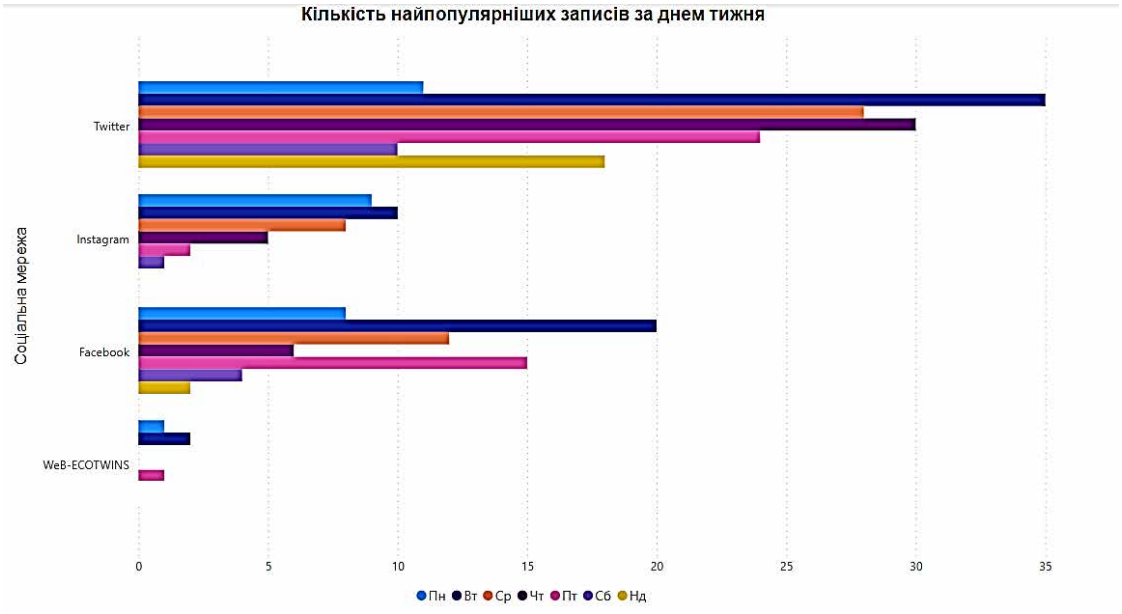


Рис. 18 Кількість найпопулярніших записів за днем тижня

В результаті ми бачимо:

- У соціальній мережі Twitter(X) найкраще публікувати твіти(пости) у вівторок та четвер.
- У соціальній мережі Instagram варто публікувати пости у понеділок, та у вівторок.

- У соціальній мережі Facebook краще публікувати новини у вівторок та середу.
- На вебсайті проекту новини варто публікувати у вівторок або п’ятницю.

На наступній діаграмі можна якісно оцінити, в якої із соціальних мереж/веб сайті є найвищий відсоток записів, що мають оцінку більше половини (рис.19). Для цієї діаграми спроектовано вибірку записів, де публікації, мають більше за середнє значення вподобань – [N ≥ (max-min)/2].

Рис. 19 Відсотковий склад популярних постів

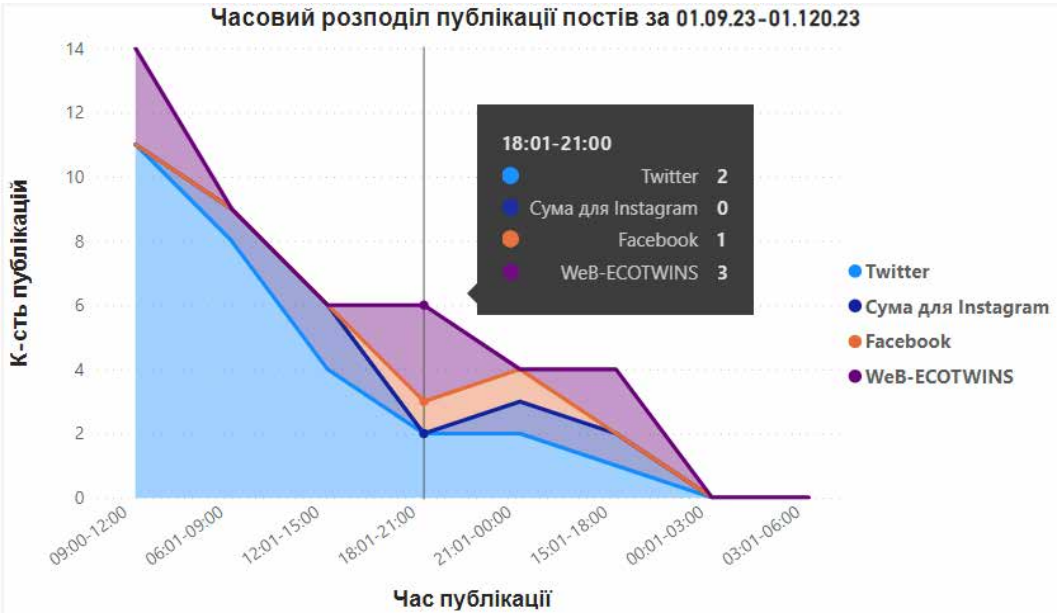


Рис. 20 Часовий розподіл публікації постів

На діаграмі,зображеній на рис. 20 можна побачити,коли і в якій соціальній мережі/вебсайті публікувались новини.

Ще один звіт розкриває історію дописів та в яких форматах вони мають найкращий фідбек серед підписників. На рисунку 21 показано які медіа дані з записів користуються найбільшою популярністю.До аналізу брались лище соцмережі, так як на вебсайті неможливо прослідкувати цей процес.

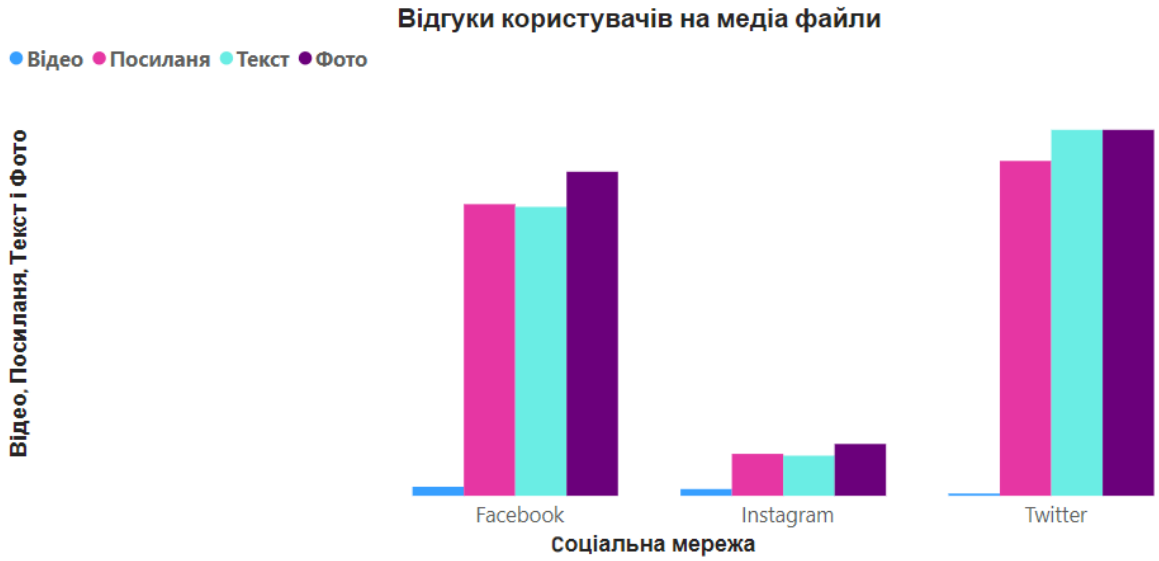


Рис. 21 Відгуки користувачів на медіа файли

На цій діаграмі було проведено аналіз по кожній сторінці в кожній соціальній мережі окремо, однак є спільна ознака - користувачі реагують на пости з фото найкраще. Також, можна помітити, що відвідувачі наукового проекту люблять і читати. Тому текстовий формат постів також має популярність серед підписників та читачів проекту. Тому, вирішено звернути увагу на кількість символів тексту в публікаціях.

Впродовж звіту про тип медіа, прийнято рішення проаналізувати,якої ж довжини пости читають відвідувачі та підписники сторінок (рис.22).

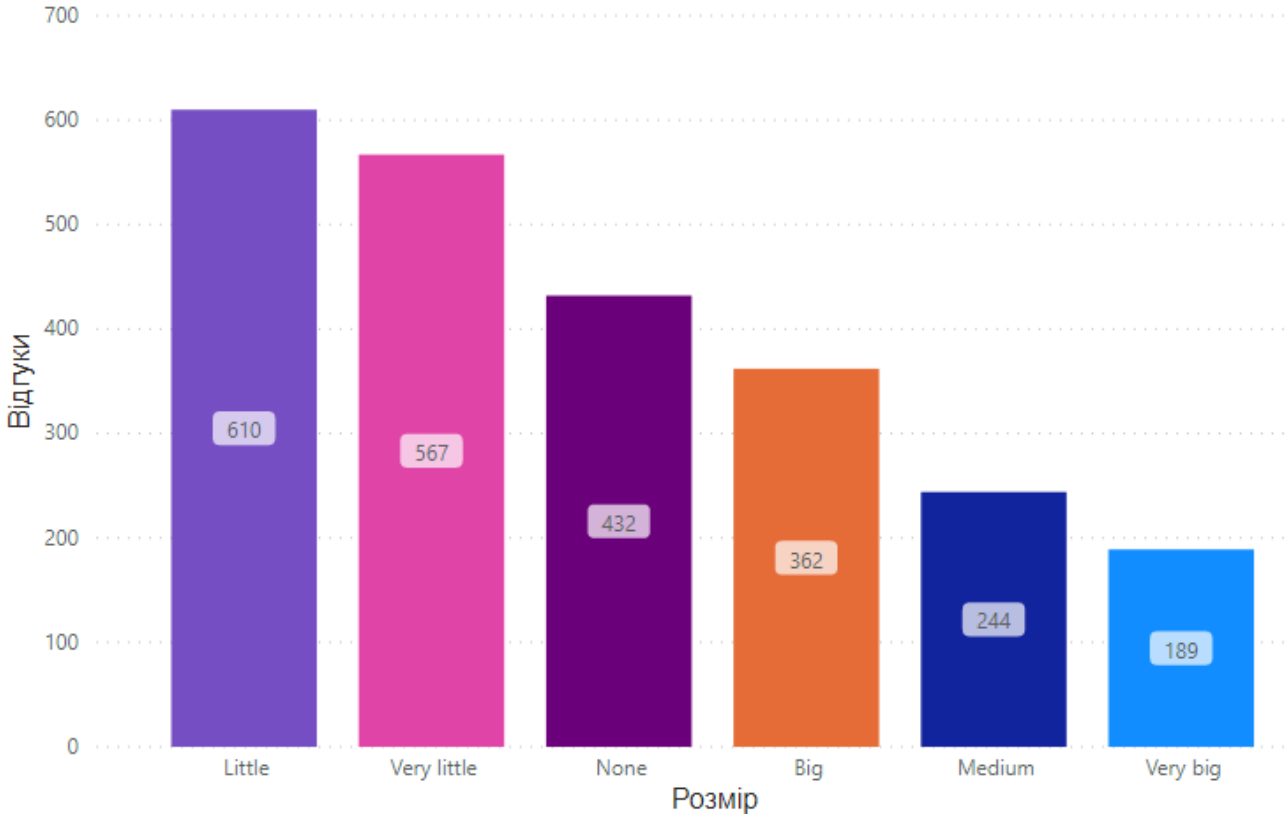


Рис. 22 Відгуки користувачів на розмір тексту

На наступній діаграмі, зображеній на рис. 23 продемонстровано кількість постів у відсотках,які мають хоча б одну позначку геолокації чи відмітку особи. Адміністратори сторінок мають розуміти, додані геолокація та відмітка осіб дають можливість охопити та зацікавити більше користувачів.

Рис. 23 Наявність геолокації та відмітки осіб в публікаціях

3.3.2 Розрахунок KPI. Ключові показники ефективності (KPI) - це важливі показники, що використовуються для вимірювання та оцінки результатів діяльності підрозділу або підприємства. Вони допомагають організації в досягненні своїх стратегічних і тактичних цілей. Використання KPI дозволяє підприємствам:

1. Оцінювати виконання цілей: KPI дозволяють підприємству визначити, наскільки успішно вони досягають своїх цілей та стратегічних завдань.
2. Моніторити продуктивність: Вони служать інструментом для постійного моніторингу продуктивності та виявлення варіацій в діяльності.
3. Покращувати прийняття рішень: KPI надають даний вихід для прийняття обґрунтованих рішень та корекції стратегії.
4. Сприяти мотивації персоналу: Вони можуть бути використані для створення системи мотивації та стимулювання працівників.
5. Оцінювати важливі аспекти діяльності: KPI можуть бути різними в залежності від конкретної стратегії компанії та її цілей.
6. Покращувати процеси і результати: Підприємства можуть вдосконалювати свої операційні процеси, спираючись на аналіз KPI.

Застосування KPI розрізняється в залежності від галузі та мети організації. Вони можуть бути використані для визначення результатів роботи управлінського персоналу, моніторингу фінансових показників, оцінки якості обслуговування клієнтів та багатьох інших аспектів діяльності. KPI допомагають компаніям зосередитися на найважливіших показниках та вдосконалювати їх для досягнення бажаних результатів.

Ключові показники ефективності можна розділити на:

1. **Запізнілі KPI:** Ці показники відображають результати діяльності після завершення визначеного періоду. Зазвичай це включає фінансові показники, такі як прибуток, витрати, прибутковість та інші фінансові

результати. Запізнілі KPI надають інформацію про те, як добре підприємство справлялося в минулому, але не завжди можуть служити для прийняття негайних дій для виправлення ситуації.

2. **Випереджаючі KPI:** Ці показники дають можливість управляти ситуацією протягом звітного періоду та спрямовувати дії на досягнення заданих результатів. Вони дозволяють спостерігати за поточною діяльністю підрозділів та компанії загалом і вчасно втручатися, якщо є потреба. Випереджаючі KPI можуть включати операційні показники, такі як обсяги виробництва, терміни доставки, рівень задоволеності клієнтів, та інші показники, які допомагають передбачити майбутні грошові потоки та якість процесів.

Впровадження ключових показників ефективності (KPI) призводить до наступних позитивних результатів:

1. Покращення продуктивності підприємства: KPI допомагають виявити та виправити слабкі місця в роботі підприємства з метою досягнення своїх стратегічних цілей.
2. Чітке визначення важливих факторів і показників: KPI роблять видимими основні аспекти, які впливають на успіх діяльності.
3. Створення планів на основі цілей: KPI допомагають розробити плани та стратегії, орієнтовані на досягнення поставлених завдань.
4. Установлення реалістичних стандартів: KPI дозволяють встановлювати реальні та досяжні нормативи для оцінки продуктивності та результатів.
5. Об'єктивна оцінка роботи персоналу: За допомогою KPI можна об'єктивно оцінювати внесок співробітників у досягнення цілей компанії.

Крім того, KPI корисні не тільки для вищого керівництва, але й для звичайних працівників, оскільки вони надають інформацію про те, як їхні дії впливають на результати та як вони можуть сприяти досягненню цілей

компанії. Таким чином, KPI стають важливим інструментом для ефективного управління та мотивації всього колективу на всіх рівнях організації [12].

У службах SQL Server Analysis Services (SSAS), ключовий індикатор продуктивності (KPI) представляє собою збір обчислень, пов'язаних із групою показників в мультидименсійному кубі. Він використовується для оцінки успішності бізнесу в рамках аналізу даних. Ключовий індикатор продуктивності визначається для обраних мір куба та включає в себе фактичні значення мір та формули для обчислення результатів продуктивності. Ці результати вимірювань показують тренд та стан продуктивності, що допомагає аналізувати та контролювати ефективність бізнес-процесів.

Для розрахунку KPI сторінок з соціальних мережта вебсайту проекту Ecotwins використано кількісні показники публікацій: коментарі, репости, перегляди та лайки.

У ході виконання було визначено такі KPI:

1. KPI_WeBEcotwins_activity
2. KPI_Twitter_activity
3. KPI_Facebook_activity
4. KPI_Instagram_activity

Кожен з показників показує активність користувачів у соцмережах у розрізі лайків та коментарів, а на вебсайті активність користувача вказується у розрізі відвідуваності та репостів.

Для розрахунку KPI необхідно розраховане фактичне значення, цільове значення та зазначені умови для визначення статусу (рис. 24-26).

Цільове значення обираємо різним для кожної зі спільнот. Так як ціль задана з самого початку замовником. Нам відомо: ціль для Фейсбуку = 60; для інстаграму = 16; для твітеру(x) = 75; для вебсайту = 65.

Аби KPI показував реальні показники сторінок в соцмережах і на вебсайті, кожній метриці надано коефіцієнт: вподобання – 0,3; коментарі – 0,3; поширення – 0,3; перегляди – 0,1.

Дані коефіцієнти ґрунтуються на цілях та інтересах адміністраторів груп. Головною метою діяльності в соціальній мережі Фейсбук в межах сторінок та відкритих груп є підвищення медіа активності читачів, відповідно підвищення метрик вподобання, коментарі та поширення. Саме тому, нами було надано коефіцієнтну перевагу даним критеріям.

Розглядаючи метрику переглядів, варто зауважити, що даний показник особливо важливий для бізнес сторінок (об'єкти даного дослідження не являються бізнес сторінками) при налаштуванні таргетованої реклами. В рамках даного дослідження кількість переглядів є найбільш невпливовим показником: підписники можуть переглянути публікацію, але ніякої взаємодії з ним не буде.

Value Expression

```
([Measures].[Num Comment], [Community Dim].[Community ID].&[Soc3]) / ([Measures].[Num
  Like], [Community Dim].[Community ID].&[Soc3])
```

No issues found Ln: 1 Ch: 133 Col: 47 SPC CRLF

Рис.24 Розрахунок фактичного значення:

Goal Expression

```
16
```

No issues found Ln: 1 Ch: 4 SPC CRLF

Рис 25. Цільове значення за замовчуванням

Status expression:

```
CASE
WHEN KPIVALUE( "KPI_Instagram_activity" ) >= KPIGOAL( "KPI_Instagram_activity" ) THEN 1
WHEN KPIVALUE( "KPI_Instagram_activity" ) > 0.1 AND KPIVALUE( "KPI_Instagram_activity" ) <
  KPIGOAL( "KPI_Instagram_activity" ) THEN 0
```

No issues found Ln: 5 Ch: 4 SPC CRLF

Рис.26 Визначення статусу за визначеними умовами

Результат розрахунку KPI зображено на рис. 27.

Display Stucture	Value	Goal	Status
KPI_Facebook_activity	58	60	
KPI_Instagram_activity	2,3	16	
KPI_WeBEcotwins_activity	12,4	75	
KPI_Twitter_activity	60	65	

Рис. 27 Розрахунок результату KPI

На основі розрахованих даних можна зробити висновок, що найбільша активність та зацікавленість користувачів в соціальних мережах Facebook та Twitter(X).

Також, можемо бачити, що показник KPI для соцмережі Instagram та вебсайту є нижчий за цільове значення. Це означає, що активність присутня, але її не достатньо для досягнення зазначених цілей.

3.3.3 Інтелектуальний аналіз даних Data Mining. Data mining – відомий як видобування даних, інтелектуальний аналіз даних або глибинний аналіз даних, представляє собою загальний термін, який використовується для опису сукупності методів, спрямованих на виявлення раніше невідомих, складних, практично використовуваних знань у наборі даних. Ці знання є важливими для прийняття рішень в різних галузях людської діяльності. Термін був вперше введений Григорієм П’ятецьким-Шапіро в 1989 році [22].

Методи Data Mining базуються на використанні різних підходів, таких як класифікація, моделювання та прогнозування, і включають в себе використання інструментів, таких як дерева рішень, штучні нейронні мережі, генетичні алгоритми, еволюційне програмування, асоціативна пам'ять і нечітка логіка. До методів Data Mining також входять статистичні підходи, такі як дескриптивний аналіз, кореляційний та регресійний аналіз, факторний аналіз, дисперсійний аналіз, компонентний аналіз, дискримінантний аналіз, аналіз часових рядів, аналіз вживаності та аналіз зв'язків.

Однією з ключових ролей методів Data Mining є можливість наглядно відображати результати обчислень через візуалізацію. Це робить інструменти Data Mining доступними для людей, які не мають спеціалізованої математичної підготовки. Використання статистичних методів аналізу даних, навпаки, вимагає глибокого розуміння теорії ймовірностей і математичної статистики.

В BI MS SQL Server створено структуру інтелектуального аналізу на основі існуючого кубу, визначено виміри вихідного кубу та вказано ключ

варіантів. На рис. 28 вказано стовпці, що використовувались в моделі інтелектуального аналізу даних, та зазначено, що стовпець «Sum Components» буде прогнозованим. На рис.29 визначено вміст та тип даних стовпців.

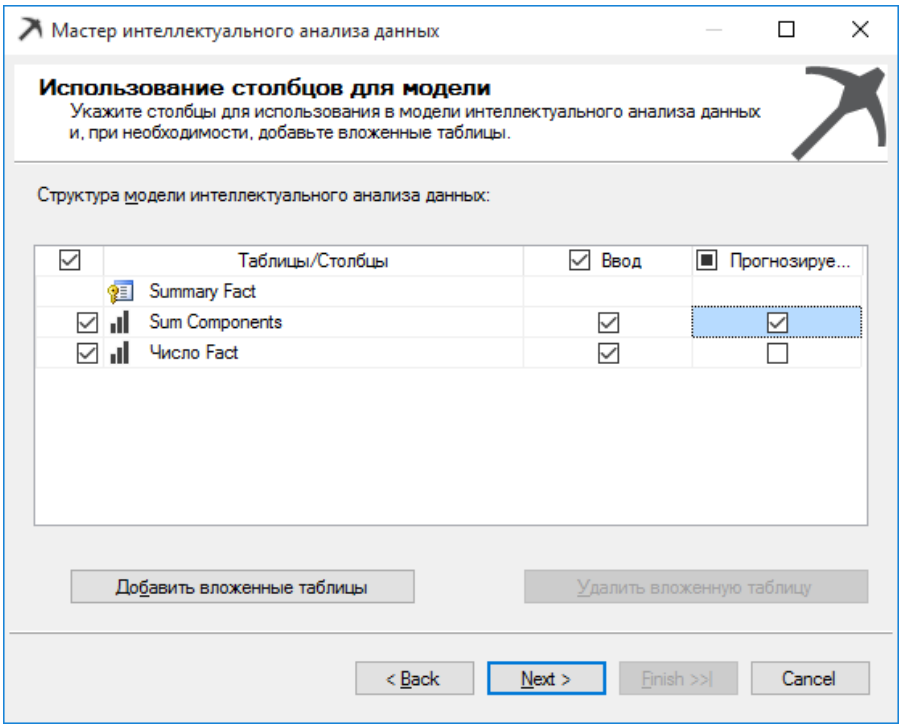


Рис. 28 Використання стовпців для моделі

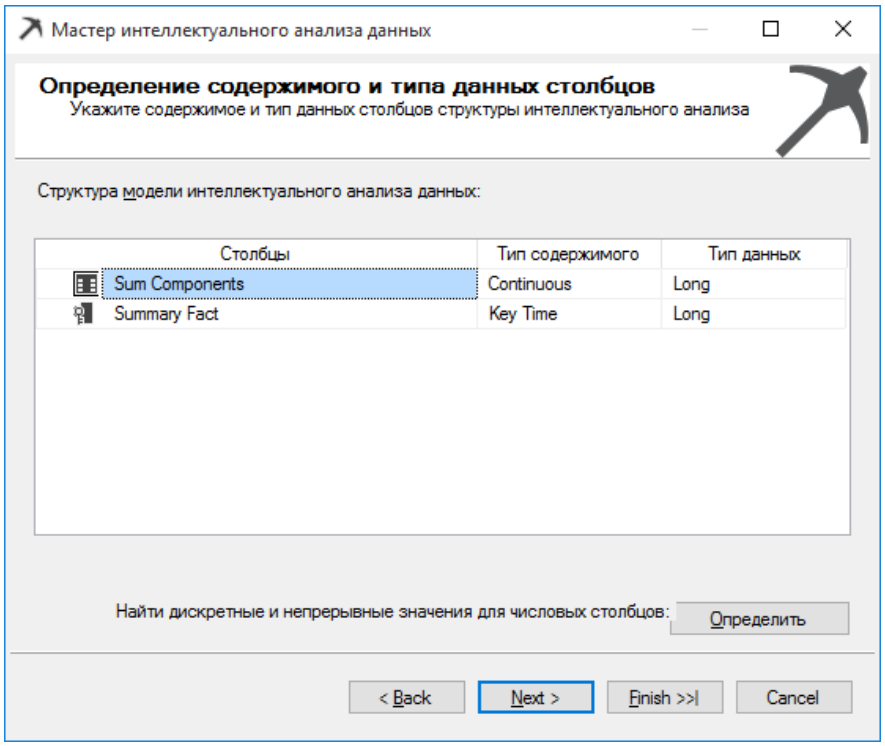


Рис. 29 Визначення вмісту та типу даних

В результаті, показано діаграму (рис. 30), де зображені значення суми компонентів (вподобання, коментарі, поширення та перегляди). Візуально

оцінивши розподіл значень, не можна виявити конкретної залежності, так як на значення «відгук відвідувача» впливає багато факторів.

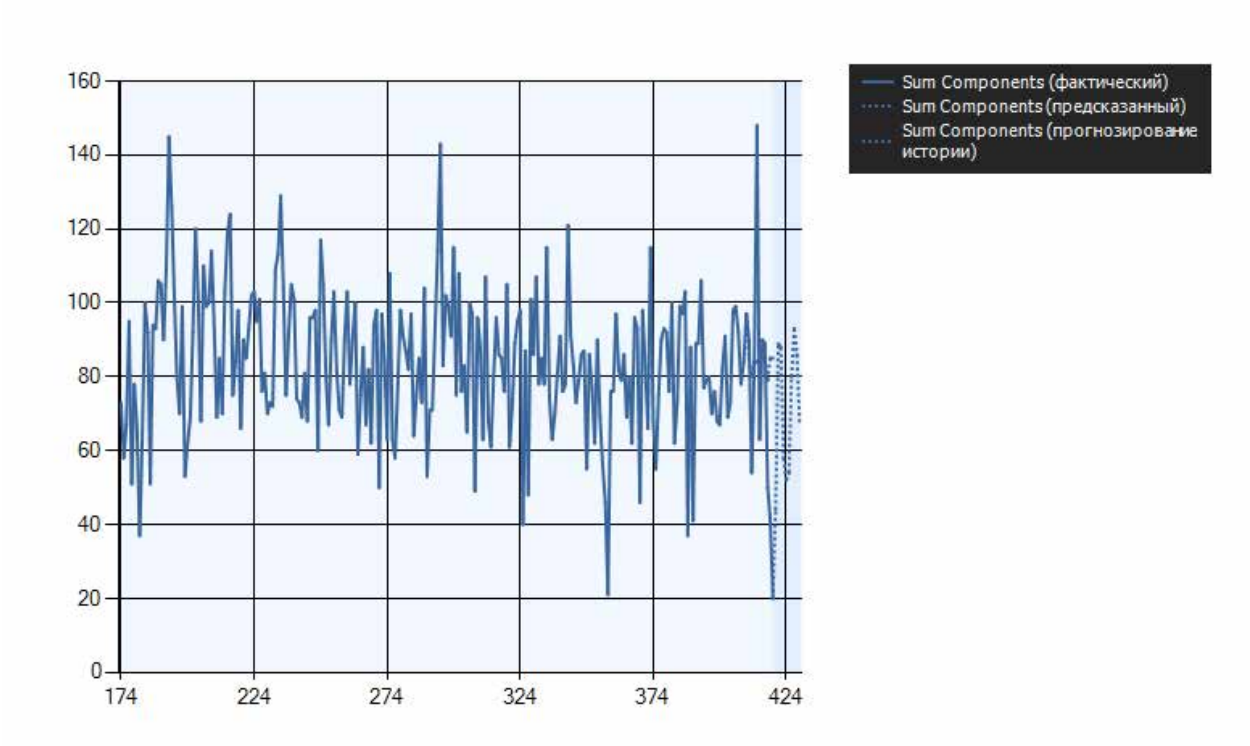


Рис. 30 Графік часового ряду

Видно, деякі пости (ID 89, 192, 294, 413) мають велику оцінку (рис.31). Переглянувши їх ознаки, можна побачити, що вони мають однаковий тип запису (мають посилання) та взагалі без тексту (Post19).

Summary_fact	Time_ID	Post_ID	Type_ID	Community_ID	Num_Lik	Num_Commen	Num_Shar	Num_Viev	Sum_components
089	time89	Post19	Type4	Com1	8	0	2	134	144
192	time192	Post19	Type4	Com1	2	2	0	141	145
294	time294	Post19	Type4	Com1	48	0	0	95	143
413	time413	Post19	Type4	Com1	85	0	0	63	148

Рис. 31 Властивості постів, що мають найвищу реакцію читачів

Далі на рисунку 32 увагу сконцентровано на прогнозованих значеннях(табл. 3).

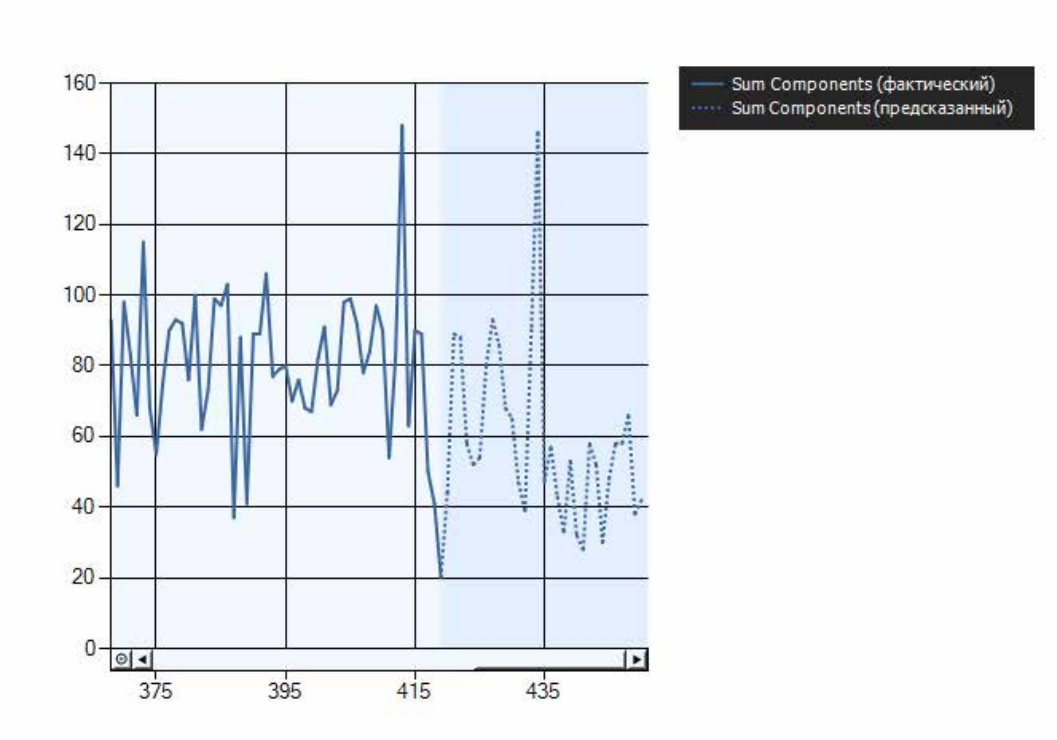


Рис. 32 Прогнозовані значення

Таблиця 3

Прогнозовані значення

Номер посту	**BI MS SQL Server**	**Номер посту**	**BI MS SQL Server**	**Номер посту**	**BI MS SQL Server**
1	**2**	**3**	**4**	**5**	**6**
419	20	429	68	439	53
420	44	430	65	440	32
1	**2**	**3**	**4**	**5**	**6**
421	89	431	47	441	28
422	88	432	39	442	58
423	58	433	91	443	52
424	52	434	146	444	30
425	54	435	47	445	48
426	80	436	57	446	58
427	93	437	43	447	58

428	86	438	33	448	66
				449	38

Наступний алгоритм Data Mining, використаний для аналізу даних це 1-Rule. OneR, скорочення від «One Rule», — це простий, але точний алгоритм класифікації, який генерує одне правило для кожного предиктора в даних, а потім вибирає правило з найменшою загальною помилкою як «одне правило».

Щоб створити правило для предиктора, ми створюємо таблицю частот для кожного предиктора щодо цілі. Було показано, що OneR створює правила лише трохи менш точні, ніж сучасні алгоритми класифікації, водночас створюючи правила, які легко інтерпретувати людиною [21].

Реалізація алгоритму 1-Rule

Для реалізації алгоритму 1-Rule було виділено 2 класи високої ("HIGH") та низької ("LOW") кількості переглядів, показник яких визначається у порівнянні з середнім показником кількості переглядів за весь період дослідження. За допомогою SQL визначені результати алгоритму. Нижче наведено реалізацію та результат у вигляді таблиці.

Спочатку проведено вибірку даних, де визначено середню кількість переглядів та клас перегляду за кожною соціальною мережею (рис. 33).

```
DECLARE @avgView FLOAT;
SELECT @avgView = AVG(Summary_Fact.Num_View)
FROM Summary_Fact;

SELECT Community_Dim.Name_of_social_community,
       Type_Dim.Type_of_post,
       Summary_Fact.Num_View AS numberOfView,
       @avgView              AS avgView,
       CASE
           WHEN @avgView > Summary_Fact.Num_View
               THEN 'LOW'
           ELSE 'HIGH'
           END               AS viewClassification
FROM Summary_Fact
         JOIN Community_Dim ON Summary_Fact.Community_ID = Community_Dim.Community_ID
         JOIN Type_Dim ON Summary_Fact.[Type_ID] = Type_Dim.[Type_ID]
```

	Name_of_social_community	Type_of_post	numberOfView	avgView	viewClassification
1	Twitter-ECOTWINS	text_photo_video	323	202	HIGH
2	Facebook-ECOTWINS	text_photo_video	242	202	HIGH
3	WeBEcotwins-ECOTWINS	text_photo	282	202	HIGH
4	Facebook-ECOTWINS	video	223	202	HIGH
5	Twitter-ECOTWINS	text_photo	119	202	LOW
6	WeBEcotwins-ECOTWINS	text_photo_video	25	202	LOW
7	Facebook-ECOTWINS	photo_video	57	202	LOW
8	Twitter-ECOTWINS	link	392	202	HIGH
9	Instagram-ECOTWINS	link	10	202	LOW
10	Twitter-ECOTWINS	photo_video	65	202	LOW
11	Facebook-ECOTWINS	text_link	279	202	HIGH
12	WeBEcotwins-ECOTWINS	text_link	223	202	HIGH
13	Twitter-ECOTWINS	text_photo_video	266	202	HIGH
14	Facebook-ECOTWINS	text_video	364	202	HIGH
15	Facebook-ECOTWINS	text	196	202	LOW
16	Twitter-ECOTWINS	video	348	202	HIGH

Рис. 33 Вибірка та класифікація фактів

Наступним етапом є визначення класифікації кількості переглядів та її ймовірність відносно середнього показника серед заданих фактів відповідно до соціальної мережі (рис.34) та типу посту (рис. 35).

```
DECLARE @avgView FLOAT;
SELECT @avgView = AVG(Summary_Fact.Num_View)
FROM Summary_Fact;

SELECT Community_Dim.Name_of_social_community,
       (SELECT COUNT(sf.Num_View)
        FROM Summary_Fact sf
        WHERE @avgView > (sf.Num_View)
          AND Summary_Fact.Community_ID = sf.Community_ID)                    AS
lessThanAvg,
       (SELECT COUNT(sf.Num_View)
        FROM Summary_Fact sf
        WHERE @avgView < (sf.Num_View)
          AND Summary_Fact.Community_ID = sf.Community_ID)                    AS
biggerThanAvg,
       COUNT(Summary_Fact.Num_View)                                           as
allView,
       AVG(Summary_Fact.Num_View)                                             as
avgPostView,
```

```
        @avgView                                                    as
avgView,
        CASE
            WHEN @avgView > AVG(Summary_Fact.Num_View)
                THEN 'LOW'
            ELSE 'HIGH'
            END                                                     AS
viewClassification,
        ROUND(CAST(CAST((SELECT COUNT(sf.Num_View)
                         FROM Summary_Fact sf
                         WHERE @avgView > (sf.Num_View)
                           AND Summary_Fact.Community_ID = sf.Community_ID) AS FLOAT) /
                    CAST(COUNT(Summary_Fact.Num_View) AS FLOAT) AS FLOAT) * 100, 2) AS
lessProbability,
        ROUND(CAST(CAST((SELECT COUNT(sf.Num_View)
                         FROM Summary_Fact sf
                         WHERE @avgView < (sf.Num_View)
                           AND Summary_Fact.Community_ID = sf.Community_ID) AS FLOAT) /
                    CAST(COUNT(Summary_Fact.Num_View) AS FLOAT) AS FLOAT) * 100, 2) AS
biggerProbability
FROM Summary_Fact
          JOIN Community_Dim ON Summary_Fact.Community_ID = Community_Dim.Community_ID
GROUP BY Community_Dim.Name_of_social_community, Summary_Fact.Community_ID
```

	Name_of_social_community	lessThanAvg	biggerThanAvg	allView	avgPostView	avgView	viewClassification	lessProbability	biggerProbability
1	Facebook-ECOTWINS	16	28	44	236	202	HIGH	36,36	63,64
2	Twitter-ECOTWINS	36	41	78	203	202	HIGH	46,15	52,56
3	WeBEcotwins-ECOTWINS	39	37	77	194	202	LOW	50,65	48,05
4	Instagram-ECOTWINS	26	16	42	169	202	LOW	61,9	38,1

Рис. 34 Класифікація відповідно до соціальної мережі

```
DECLARE @avgView FLOAT;
SELECT @avgView = AVG(Summary_Fact.Num_View)
FROM Summary_Fact;

SELECT Type_Dim.Type_of_post,
       (SELECT COUNT(sf.Num_View)
        FROM Summary_Fact sf
        WHERE @avgView > (sf.Num_View)
          AND Summary_Fact.[Type_ID] = sf.[Type_ID])                AS lessThanAvg,
       (SELECT COUNT(sf.Num_View)
        FROM Summary_Fact sf
        WHERE @avgView < (sf.Num_View)
          AND Summary_Fact.[Type_ID] = sf.[Type_ID])                AS biggerThanAvg,
```

```
        COUNT(Summary_Fact.Num_View)                                    as
allView,
        AVG(Summary_Fact.Num_View)                                      as
avgPostView,
        @avgView                                                        as
avgView,
        CASE
            WHEN @avgView > AVG(Summary_Fact.Num_View)
                THEN 'LOW'
            ELSE 'HIGH'
            END                                                         AS
viewClassification,
        ROUND(CAST(CAST((SELECT COUNT(sf.Num_View)
                         FROM Summary_Fact sf
                         WHERE @avgView > (sf.Num_View)
                           AND Summary_Fact.[Type_ID] = sf.[Type_ID]) AS FLOAT) /
                   CAST(COUNT(Summary_Fact.Num_View) AS FLOAT) AS FLOAT) * 100, 2) AS
lessProbability,
        ROUND(CAST(CAST((SELECT COUNT(sf.Num_View)
                         FROM Summary_Fact sf
                         WHERE @avgView < (sf.Num_View)
                           AND Summary_Fact.[Type_ID] = sf.[Type_ID]) AS FLOAT) /
                   CAST(COUNT(Summary_Fact.Num_View) AS FLOAT) AS FLOAT) * 100, 2) AS
biggerProbability
FROM Summary_Fact
         JOIN Type_Dim ON Summary_Fact.[Type_ID] = Type_Dim.[Type_ID]
                  GROUP BY Type_Dim.Type_of_post, Summary_Fact.[Type_ID]
```

	Type_of_post	lessThanAvg	biggerThanAvg	allView	avgPostView	avgView	viewClassification	lessProbability	biggerProbability
1	text	7	6	14	206	202	HIGH	50	42,86
2	photo	13	14	27	215	202	HIGH	48,15	51,85
3	video	16	16	32	183	202	LOW	50	50
4	link	12	23	35	230	202	HIGH	34,29	65,71
5	text_photo	14	14	28	204	202	HIGH	50	50
6	text_video	12	13	25	215	202	HIGH	48	52
7	text_link	16	14	31	212	202	HIGH	51,61	45,16
8	photo_video	12	12	24	191	202	LOW	50	50
9	photo_link	14	11	25	179	202	LOW	56	44
10	video_link	22	11	33	174	202	LOW	66,67	33,33
11	text_photo_video	11	9	20	181	202	LOW	55	45
12	text_photo_link	7	9	16	245	202	HIGH	43,75	56,25

Рис. 35 Класифікація відповідно до типу посту

Аналізуючи отримані дані можна виділити наступні правила:

- Якщо соціальна мережа Facebook, то клас перегляду високий ("High") з імовірністю 63,64%.

- Якщо вебсайт WeBEcotwins, то клас перегляду низький ("Low") з імовірністю 48,05%.
- Якщо тип посту посилання, то клас перегляду високий ("High") з імовірністю 65,71%.
- Якщо тип посту текст, фото та посилання, то клас перегляду високий ("High") з імовірністю 56,25%.

Можна зробити висновок, що на класифікацію і кількість переглядів впливає соціальна мережа, у якій поширюється інформація.

Наступним методом класифікації, який ми розглянемо, є метод наївного Байєса. Цей метод базується на теоремі Байєса і передбачає, що ознаки є незалежними одна від одної в межах класу. Іншими словами, наївний Байєсів класифікатор припускає, що наявність певної ознаки в класі не має взаємозв'язку з наявністю інших ознак.

Наприклад, в методі наївного Байєса фрукт можна вважати яблуком, якщо він володіє такими характеристиками, як червоний колір, кругла форма і діаметр приблизно 3 дюйми. Навіть якщо ці характеристики можуть бути взаємозалежними або залежати від наявності інших факторів, наївна модель Байєса припускає, що ці характеристики незалежно впливають на ймовірність того, що фрукт є яблуком. Тому вона отримала назву "наївною".

Модель наївного Байєса легко будується і особливо корисна для аналізу великих обсягів даних. Незважаючи на свою простоту, наївна модель Байєса відома своєю здатністю перевершувати навіть складніші методи класифікації.

Теорема Байєса забезпечує спосіб обчислення апостеріорної ймовірності P(c|x) з P(c), P(x) і P(x|c). Розберемо рівняння,яке зображено на Рис.31

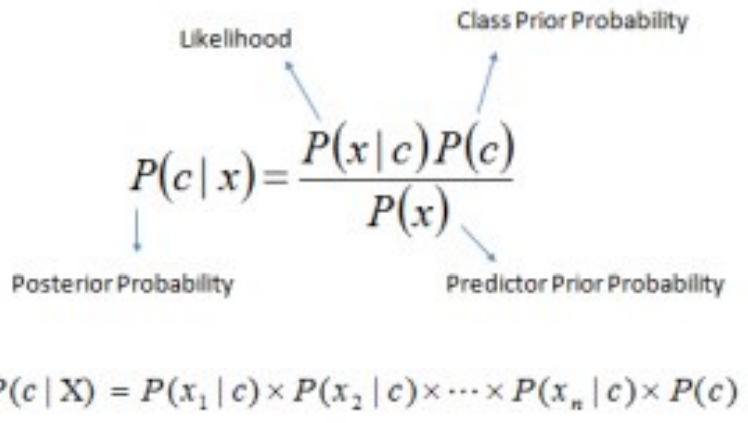


Рис.31 Апостеріорна ймовірність

Вище,

P(c|x) — апостеріорна ймовірність класу (c, ціль), заданого предиктором (x, атрибути).

P(c) — попередня ймовірність класу.

P(x|c) – це ймовірність, яка є ймовірністю провісника даного класу.

P(x) – попередня ймовірність предиктора.

Плюси:

- Він легко та швидко передбачає клас тестового набору даних і вдало використовується у завданнях багатокласового прогнозування.
- Коли виконуються припущення про незалежність між ознаками, наївний байєсовий класифікатор може працювати краще, порівняно з іншими моделями, такими як логістична регресія, і вимагає менше навчальних даних.
- Він ефективно застосовується до категорійних вхідних змінних у порівнянні з числовими змінними. Для числових змінних передбачається нормальний розподіл, хоча це може бути сильним припущенням.

Мінуси:

- Якщо категоріальна змінна має категорію в тестовому наборі даних, яка не мала відповідника в навчальному наборі, модель призведе до

нульової ймовірності, що може призвести до некоректної класифікації. Це часто зветься "нульовою частотою". Для вирішення такої проблеми часто застосовують техніку згладжування, і одним з простих методів згладжування є оцінка Лапласа.

- Наївний Байєс іноді вважається ненадійним оцінювачем, тому результати ймовірностей, які надаються через метод predict_proba, не завжди є абсолютно точними.
- Ще одним обмеженням наївного Байєса є припущення про незалежність між предикторами. У реальних даних досить рідко можна знайти абсолютно незалежні предиктори.

Застосування наївних байєсових алгоритмів включає в себе наступні області:

- Прогнозування в режимі реального часу: Наївний Байєс - це швидкий класифікатор, і його можна успішно використовувати для прогнозування в режимі реального часу.
- Багатокласове передбачення: Цей алгоритм також відомий своєю здатністю передбачати ймовірність кількох класів цільової змінної, що робить його ефективним для задач багатокласової класифікації.
- Класифікація тексту / Фільтрація спаму / Аналіз настроїв: Наївні байєсівські класифікатори особливо ефективні в класифікації тексту завдяки їхній здатності до працювати з багатьма класами і на підставі припущення про незалежність. Вони широко використовуються в фільтрації спаму (визначення спаму в електронній пошті) та аналізі настроїв (в соціальних мережах для визначення позитивних та негативних настроїв користувачів).
- Системи рекомендацій: Наївний Байєс разом із спільною фільтрацією використовуються для створення систем рекомендацій. Вони використовують методи машинного навчання та аналізу даних для

фільтрації прихованих інформаційних ресурсів і передбачення, чи цікавий користувачеві певний ресурс [22].

Для реалізації алгоритму Naive Bayes на основі розгорнутого кубу, використовуючи службу MS SSAS, створена нова структура інтелектуального аналізу даних. Результатом проведення аналізу було сформовано мережу залежностей (рис. 36).

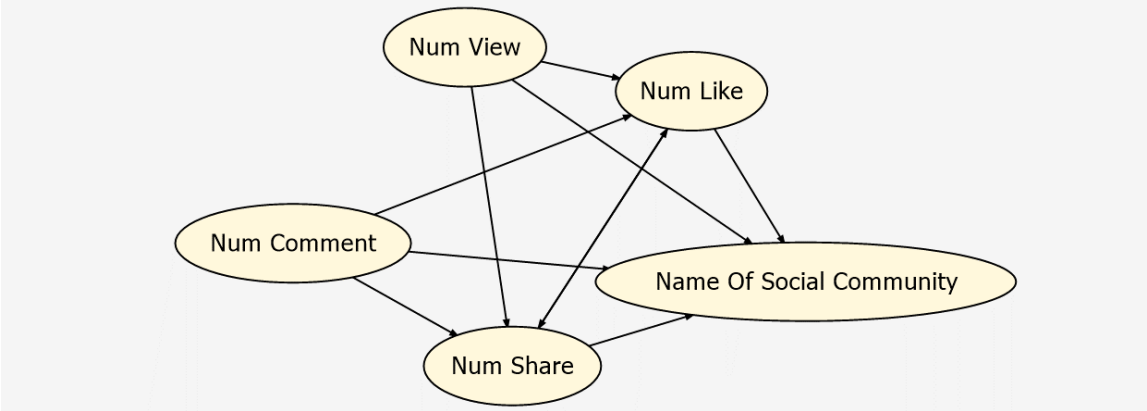


Рис. 36 Розгорнута мережа залежностей

На основі побудованої мережі можна зробити висновок, що кількість лайків, коментарів та поширень напряму корелює з кількістю переглядів. Це логічне спостереження, оскільки чим більше людей переглядають публікацію, тим більше можливостей отримати реакції від аудиторії.

На додачу, розглянемо більш детальний опис за кожним прогнозованим значенням. На рис. 37 наведено детальний опис профілю атрибуту «Name of social community». Завдяки проаналізованим даним робимо висновок, що значно впливає кількість переглядів.

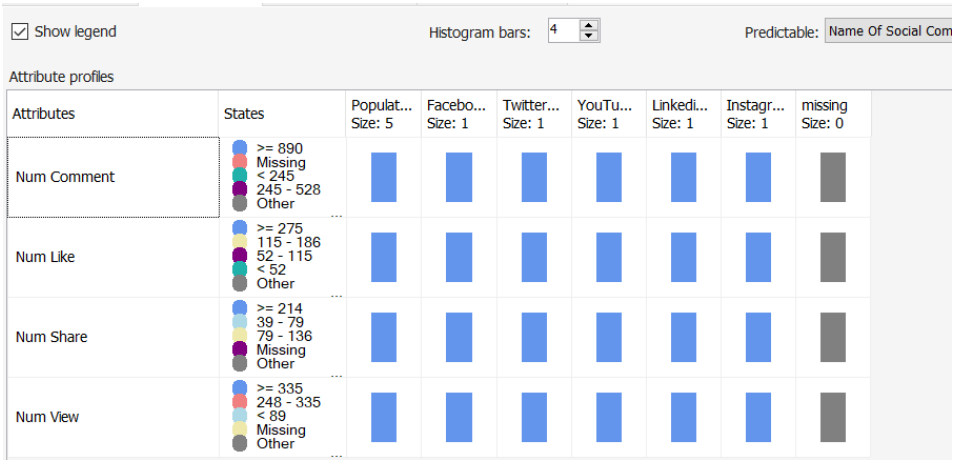


Рис 37 Профіль атрибуту «Name of social community»

Наступний алгоритм,який буде використано-це алгоритм правил-асоціацій.

Правила асоціації - це "якщо-то" оператори, які допомагають визначити ймовірність зв'язків між елементами даних у великих наборах даних у різних типах баз даних. Видобуток правил асоціацій має різноманітні застосування і широко використовується для виявлення кореляцій продажів у транзакційних даних або в наборах медичних даних.

У науці про дані, правила асоціації використовуються для пошуку кореляцій та спільного виникнення між наборами даних. Вони ідеально підходять для розкриття закономірностей у даних, які здаються незалежними, такими як реляційні бази даних і бази даних транзакцій. Процес використання правил асоціації іноді називають "видобуток правил асоціації" або "асоціаційний аналіз".

Вироблення правил асоціацій на базовому рівні включає в себе використання моделей машинного навчання для аналізу даних з метою виявлення шаблонів або супутніх випадків у базі даних. В процесі вироблення правил асоціацій визначаються часті асоціації "якщо-то", які в самому собі є правилами асоціацій.

Правило асоціації складається з двох частин: антецедент (у випадку) і консеквент (тоді). Антецедент - це елемент, знайдений у даних, і консеквент - це елемент, який співвідноситься з антецедентом у конкретних умовах. Такі правила допомагають виявити залежності та співвідношення між різними елементами у наборі даних.

Правила асоціації створюються шляхом аналізу даних для виявлення поширених "якщо-то" моделей та використання критеріїв підтримки та впевненості для ідентифікації значущих зв'язків. Підтримка вказує на те, наскільки часто певні елементи зустрічаються у наборі даних. Впевненість визначає, наскільки впевнено можна стверджувати, що "якщо-то" визнається істинним. Третій показник, називаний підвищенням, дозволяє порівнювати

впевненість з очікуваною впевненістю або тим, скільки разів очікується, що "якщо-то" буде істинним.

Правила асоціації розраховуються на основі наборів елементів, що складаються з двох чи більше елементів. Кількість можливих правил може бути дуже великою, що призводить до надмірної складності. Тому правила асоціації зазвичай створюються на основі даних, де зв'язки між елементами добре представлені і можуть бути інтерпретовані [4].

Вимірювання ефективності правил асоціації:

Сила даного правила асоціації вимірюється за допомогою двох основних параметрів: підтримки та впевненості. Підтримка вказує на частоту, з якою дане правило виявляється в видобутих даних. Впевненість визначає, наскільки часто це правило виявляється істинним у практиці. Правило може мати сильну кореляцію в наборі даних, оскільки воно зустрічається дуже часто, але може мало виконуватися, коли його застосовують. В такому випадку ми спостерігаємо високу підтримку, але низьку впевненість.

З іншого боку, правило може не бути дуже очевидним у наборі даних, але подальший аналіз може показати, що воно відповідає досить часто. У цьому випадку ми маємо високу впевненість, але низьку підтримку. Використання цих показників допомагає аналітикам відрізняти причинно-наслідкові зв'язки від кореляції і дозволяє їм належним чином оцінювати дане правило.

Третій показник, відомий як значення підйому, відображає відношення між впевненістю і підтримкою. У разі, якщо значення підйому є від'ємним, це вказує на наявність негативної кореляції між точками даних. Якщо воно додатнє, то існує позитивна кореляція, а коли відношення дорівнює 1, це свідчить про відсутність кореляції [23].

На основі розгорнутої структури було побудовано мережу залежностей назви соціальної мережі та кількості переглядів від сумарної кількості лайків, коментарів та поширень (рис. 38).

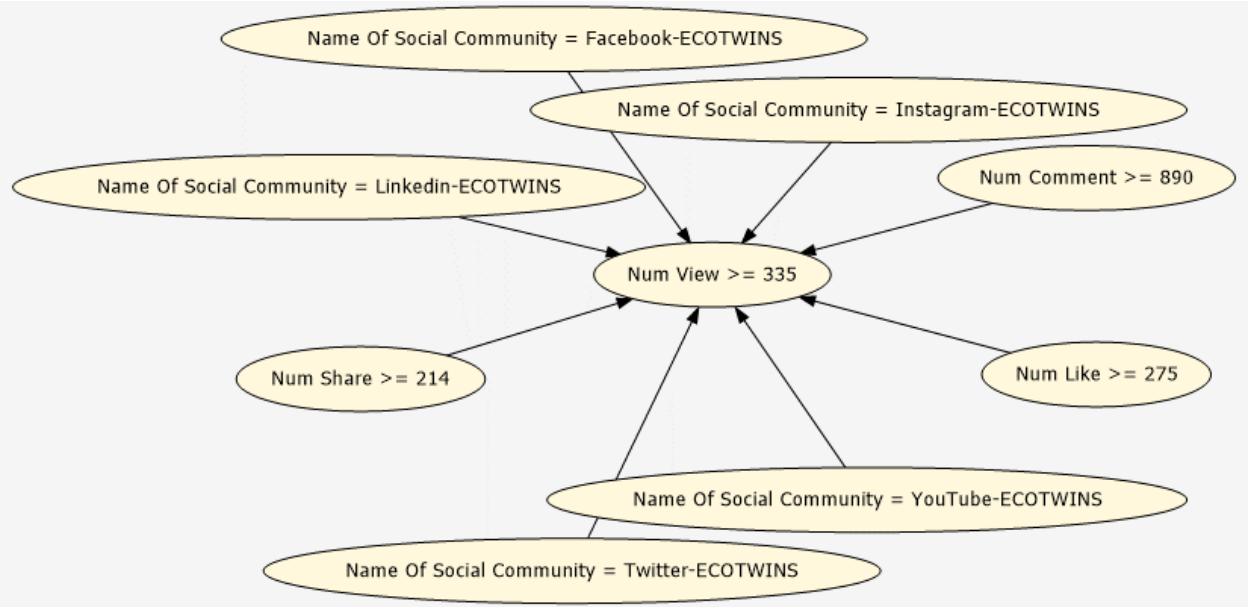


Рис. 38 Мережа залежностей

Додатково побудовано правила з вказаною імовірністю та важливістю. Деякі наведені на рисунку 39.

Probability	Importance	Rule
1,000	0,234	Num Share >= 214 -> Num View >= 335
1,000	0,234	Num Like >= 275 -> Num View >= 335
1,000	0,234	Num Comment >= 890 -> Num View >= 335
1,000	0,234	Num Share >= 214, Num Comment >= 890 -> Num View >= 335
1,000	0,234	Num Like >= 275, Num Comment >= 890 -> Num View >= 335
1,000	0,234	Num Share >= 214, Num Like >= 275 -> Num View >= 335
1,000	-0,097	Name Of Social Community = Linkedin-ECOTWINS, Num Comment >= 890 -> Num View >= 335
1,000	-0,097	Name Of Social Community = YouTube-ECOTWINS, Num Share >= 214 -> Num View >= 335
1,000	-0,097	Name Of Social Community = YouTube-ECOTWINS, Num Like >= 275 -> Num View >= 335
1,000	-0,097	Name Of Social Community = YouTube-ECOTWINS, Num Comment >= 890 -> Num View >= 335
1,000	-0,097	Name Of Social Community = Twitter-ECOTWINS -> Num View >= 335
1,000	-0,097	Name Of Social Community = Twitter-ECOTWINS, Num Share >= 214 -> Num View >= 335
1,000	-0,097	Name Of Social Community = Twitter-ECOTWINS, Num Like >= 275 -> Num View >= 335

Рис. 39 Сформовані правила на основі знайдених взаємозалежностей

У вкладці «Mining accuracy cHart» спостерігаємо за графіком, що показує відношення правильних до загальних популяцій. Також модель порівнюється з так званою «ідеальною моделлю». У нашому випадку відсоток правильних 40,48%, а в ідеальній моделі він дорівнює 50% (рис.40-41).

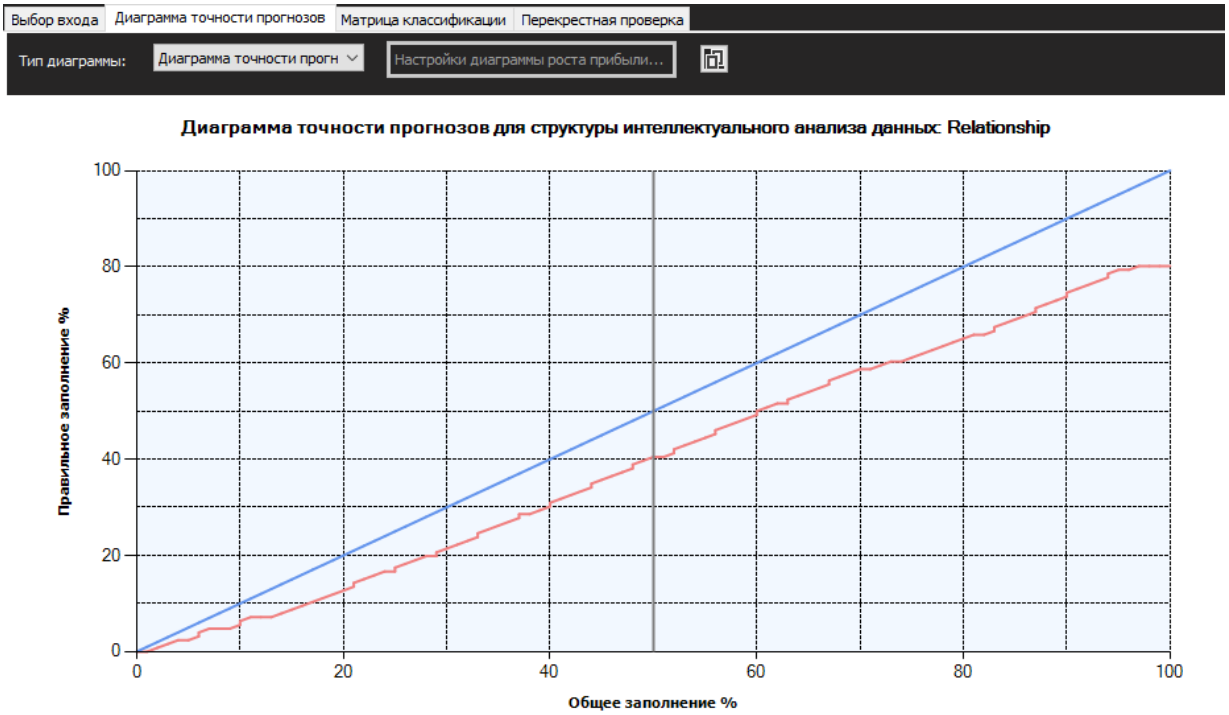


Рис. 40 Графік порівняння моделі з ідеальною моделлю

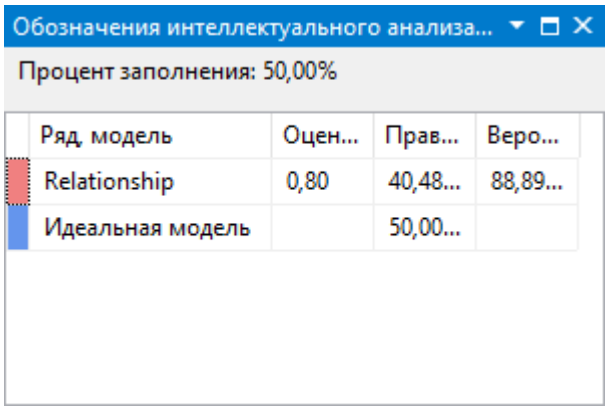
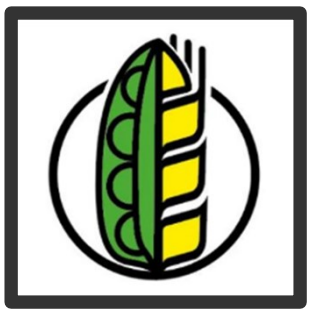


Рис. 41 Порівняння моделі з ідеальною моделлю

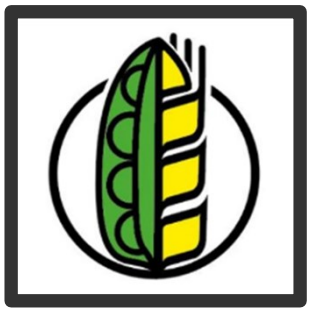
3.4 Рекомендації

Після аналізу зібраних даних був створений інформаційний додаток із рекомендаціями для адміністраторів сторінок у соціальних мережах:Twitter(X), Facebook, Instagram та на вебсайті проекту Ecotwins (Додаток Г).

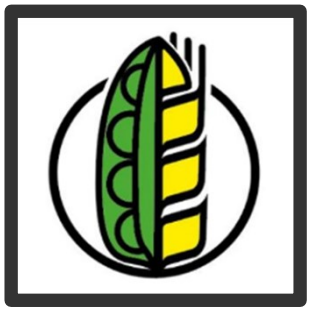
Для адміністратора кожної соціальної мережі встановлено правила та поради для управління сторінками та підвищенням активності.



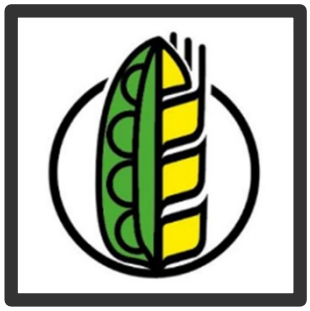
У соцмережі Facebook на сторінці проекту: «ECOTWINS-Project about researching» найкращий час для публікацій – вівторок з 12 до 15 годині дня та п’ятниця з 21 до 12 години ночі. Якщо у постах публікувати фотографії – це підвищить рівень взаємодії у три рази, а якщо відмічати людей у цих постах, то підніметься активність підписників.



У соціальній мережі Twitter(X) на сторінці проекту: «ECOTWINS-Project about researching» краще робити твіти: вівторок, середа, четвер, з 9 до 12 години. Найоптимальніший розмір постів від 100 до 300 символів, відмітки осіб привернуть увагу цільової аудиторії.



У соцмережі Instagram на сторінці проекту: «ECOTWINS-Project about researching» краще робити пости у вівторок, п’ятницю та середу з 9 по 12 години. Пости мають включати фото та посилання, а тексту має бути 350 символів, також варто відмічати геолокацію.



На вебсайті проекту: «ECOTWINS-Project about researching» найкращий час для публікацій новин це вівторок та п’ятниця з 18 по 21 годину.Так як на сайті новини складаються з фото та тексту, всі пости мають включати фото відповідної тематики новини та текст який складається не менше ніж 500 символів, відмітки посилань на соціальні мережі проекту чи учасників новин збільшать охоплення на сторінках у соціальних мережах, та збільшать інтерес до відвідуваності сайту, завдяки розповсюдженню новини на сторінках інших осіб.

Під час перевірки кількісних показників повторно було виявлено, що впровадження рекомендацій призвело до підвищення рівня активності користувачів та їх взаємодії з дописами на 62%.

ВИСНОВКИ

Під час виконання магістерської дипломної роботи було розроблено та створено базу даних на основі реляційної системи для аналізу активності користувачів на сторінках у соціальних мережах:Twitter(X), Facebook, Instagram та на вебсайті наукового проекту Ecotwins.

Для роботи були використані наступні інструменти:

- Facebook Business Suit (адмін панель керуванням сторінки проекту та вивантаження даних);
- Twitter API (для аналізу даних та їх вивантаження);
- Google Analytics (для аналітики відвідувачів вебсайту та вивантаження даних для подальшого аналізу)
- MS SQL Server 2017 (створення БД та СД);
- SQL Server Data Tools (BI) (служби SSAS та SSIS та розрахунок KPI);
- Power BI Desktop (створення звітів)

Всі дані перенесено із бази даних до новоствореного сховища даних у системі керування базами даних MS SQL Server, яке має структуру "зірка". Визначено структуру сховища даних, створено OLAP-куб, налаштовано процеси завантаження та заповнення куба за допомогою Data Flow в середовищі SQL Server BI. Далі обчислили ключові показники ефективності (KPI) та побудували звіти в середовищі Microsoft Power BI. Також провели аналіз даних методами Data Mining. Наголошую, що дані є актуальними, станом на жовтень 2023 року.

Звіти відповідають на всі поставлені питання у постановці завдання. Обчислений KPI показав як розвиваються спільноти наукового проекту в соціальних мережах:Facebook, Twitter(X), Instagram та на вебсайті Ecotwins.

За результатами дослідження створено інформаційну довідку для адміністраторів сторінок у соціальних мережах та для контент-мейкерів вебсайту (Додаток Г).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Голуб Б.Л., Ящук Д.Ю. Організація сховища даних: навч. посіб. Київ: Національний університет біоресурсів і природокористування України, 2018.

2. Голуб Б.Л., Ящук Д.Ю. Методичні вказівки до виконання курсового проекту з дисципліни «Організація сховища даних» Київ: Національний університет біоресурсів і природокористування України, 2018.

3. Сховища та простори даних: монографія / Н. Б. Шаховська, В. В. Пасічник ; М-во освіти і науки України, Нац. ун-т «Львів. політехніка». − Л. : Вид-во Нац. ун-ту «Львів. політехніка», 2009. − 240 с. − Бібліогр. : с. 230−240 (207 назв). − ISBN 978-966-553-796-0.

4. Правила асоціації. – [Електронний ресурс]. – Режим доступу: http://mmsa.kpi.ua/sites/default/files/disciplines/%D0%A0%D0%BE%D0%B7%D1%80%D0%BE%D0%B1%D0%BA%D0%B0%20%D1%96%20%D1%82%D0%B5%D1%81%D1%82%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%20%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC/didkovska_m_v_testing_lecture_5.pdf

5. О.Є. Коваленко Стандартизація формального опису системної архітектури ситуаційних центрів: Інститут проблем математичних машин і систем НАН України, м. Київ, 2015. 4 с. Режим доступу - Http://conf.atsukr.org.ua/conf_files/conf_dir_24/Kovalenko_sppr2015.pdf

6. SilberscHatz, AbraHam; KortH, Henry F.; SudarsHan, S. (2011). Database system concepts (вид. 6). New York: McGraw-Hill. ISBN 978-0-07-352332-3. OCLC 436031093. Режим доступу - Https://www.worldcat.org/oclc/436031093

7. Сховища та простори даних: монографія / Н. Б. Шаховська, В. В. Пасічник ; М-во освіти і науки України, Нац. ун-т «Львів. політехніка». − Л. : Вид-во Нац. ун-ту «Львів. політехніка», 2009. − 240 с. − Бібліогр. : с. 230−240 (207 назв). − ISBN 978-966-553-796-0.

8. Створення сховищ даних. Технології OLAP та Data Mining [Електронний ресурс]. – Режим доступу: Https://pidrucHniki.com/16120414/informatika/stvorennya_sHoviscH_daniH_teHnologiyi_olap_data_mining

9. .Опис BI. – [Електронний ресурс]. – Режим доступу: https://powerbi.microsoft.com/ru-ru/what-is-power-bi/

10. SQL Servier Business intelligence [Електронний ресурс]. – Режим доступу: Https://www.microsoft.com/ru-ru/sql-server/sql-business-intelligence

11. Мюллер Р.Дж. Базы данных и UML. Проектирование [Текст] / Р.Дж. Мюллер - М.: ЛОРИ, - 2002. – 420с.

12. Визуальные элементы ключевого показателя эффективности (КПЭ) / Olprod // GitHub. – 30.01.2020 Режим доступу - Https://docs.microsoft.com/ru-ru/power-bi/visuals/power-bi-visualization-kpi

13. Політика використання даних Facebook, дата останньої редакції: 19 квітня 2018 року. Режим доступу - Https://www.facebook.com/about/privacy/update

14. Introducing Facebook Business Suite / Facebook for business // Facebook Business. – 17.09.2020. Режим доступу - Https://www.facebook.com/business/news/introducing-facebook-business-suite?utm_content=buffer4d224&utm_medium=social&utm_source=twitter&utm_campaign=buffer

15. Базы данных: проектирование / Стружкин Н.П.. - 2017. – 315с. Режим доступу: Https://studme.org/77189/informatika/bazy_dannyH_proektirovanie

16. Google Analytics. Режим доступу: https://marketingplatform.google.com/about/analytics/

17. Robert Wrembel, CHristian Koncilia. Data wareHouses and OLAP: concepts, arcHitectures, and solutions. / Wrembel R., Koncilia CH. // IRM Press. - 2007. PP. 1-26. Режим доступу - Https://books.google.ru/books?id=XFivorxZDm8C&printsec=frontcover&dq=%22Data+wareHouses+and+OLAP:+concepts,+arcHitectures,+and+solutions%22&source=bl&ots=Ss83EvzwXx&sig=qJxPFxXYo7fqKAfUk3bBiMaE_Z0&Hl=ru&ei=vrSTS5qeBpXGnAOLmt2XCw&sa=X&oi=book_result&ct=result#v=onepage&q=&f=false

18. Кращі сервіси для веб-аналітики. Режим доступу - https://www.ukraine.com.ua/uk/blog/web-analytics/luchshie-servisi-dlya-veb-analitiki.html

19. Сервіси для аналітики соцмереж. Режим доступу - https://bazilik.media/12-servisiv-dlia-analityky-sotsmerezh/

20. Ian H. Witten, Eibe Frank and Mark A. Hall. Data Mining: Practical MacHine Learning Tools and TecHniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664. — ISBN 9780123748560. Режим доступу - Https://www.elsevier.com/books/data-mining-practical-macHine-learning-tools-and-tecHniques/witten/978-0-12-374856-0

21. OneR. – [Електронний ресурс]. – Режим доступу: https://www.saedsayad.com/oner.htm

22. Six Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. – [Електронний ресурс]. – Режим доступу: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

23. What are Association Rules in Data Mining (Association Rule Mining)? – [Електронний ресурс]. – Режим доступу: https://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining\

24. Twitter API. – [Електронний ресурс]. – Режим доступу:

https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api\

25. Вебсайт Ecotwins. – [Електронний ресурс]. – Режим доступу: https://ecotwins.eu/en

26. Проектування інформаційних систем. – [Електронний ресурс]. – Режим доступу: https://eprints.cdu.edu.ua/1481/1/pro.pdf

27. Аналіз обробки даних. – [Електронний ресурс]. – Режим доступу: https://core.ac.uk/download/pdf/159817923.pdf