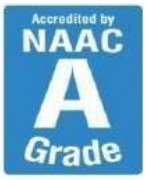
SAVEETHA

INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

(Declared as Deemed to be University under Section 3 of UGC Act 1956)



Social Media User Segmentation in Data Warehousing

CAPSTONE PROJECT REPORT

CSA1674

Data Warehousing and Data Mining for Search Engines

Submitted by

V.Nagarjuna – 192210300

Department of Computer Science and Engineering

Guided by

Dr. Porkodisivaram

Department of Computer Science and Engineering

Saveetha School of Engineering.

ABSTRACT

Social media user segmentation in data warehousing involves categorizing users into distinct segments based on their behaviors, preferences, and interactions on social media platforms. This project aims to develop a machine learning model that segments users effectively to facilitate targeted marketing, personalized content delivery, and enhanced user experience. The project addresses challenges such as data preprocessing, feature extraction, and the implementation of clustering algorithms like k-Means or hierarchical clustering. The ultimate goal is to provide a robust tool for businesses to understand and engage with their social media audience more effectively.

KEYWORDS: Social Media, User Segmentation, Data Warehousing, Clustering, k-Means

INTRODUCTION

Preliminary Stage

Assignment Description

Social media user segmentation in data warehousing involves preprocessing social media data to remove noise and standardize the data, extracting relevant features such as user demographics, behaviors, and interactions, and applying clustering algorithms to group users into distinct segments. The performance of the model is assessed using metrics such as silhouette score and cluster purity. By segmenting users, businesses can gain insights into user preferences and behaviors, enabling more targeted marketing strategies and personalized content delivery.

Assignment Work Distribution

Project Scope Definition:

The project aims to create a framework for analyzing social media users and segmenting them into meaningful groups based on their activities, preferences, and interactions. This involves gathering data from various social media platforms, preprocessing the data to ensure consistency and accuracy, extracting features relevant to user segmentation, and applying clustering algorithms to identify distinct user segments.

Data Collection and Preparation:

Identify the Data Sources:

Social media platforms such as Twitter, Facebook, and Instagram provide diverse data on user behaviors and interactions.

Develop a Data Collection Plan:

Data is collected through APIs and web scraping, supplemented by manual collection methods for user profiles and activity logs.

Data Preprocessing:

The collected data undergoes cleaning to remove noise, tokenization, normalization, and feature extraction to prepare it for clustering analysis.

Consistency:

Ensuring consistent data representation through preprocessing steps such as normalization and standardization.

Exploratory Data Analysis

Performing exploratory data analysis (EDA) involves assessing the distribution of user attributes, analyzing activity patterns, and identifying prevalent behaviors. Visualization techniques like user activity heatmaps and interaction networks help uncover insights crucial for effective user segmentation.

PROBLEM STATEMENT

The problem statement revolves around building a robust data warehousing solution capable of accurately segmenting social media users based on their behaviors and preferences. This includes preprocessing data to remove noise, extracting meaningful features, and applying clustering algorithms like k-Means to identify distinct user segments. Challenges include handling diverse data types, ensuring scalability, and maintaining model performance over time.

PROPOSED DESIGN WORKS

Data Collection

Data collection involves gathering user data from various social media platforms using APIs and web scraping techniques. This includes user demographics, activity logs, and interaction data.

Data Processing

Data preprocessing involves cleaning the data, normalizing it, and extracting relevant features such as user engagement metrics, activity patterns, and interaction frequencies.

Clustering Algorithms

Clustering algorithms like k-Means and hierarchical clustering are employed to segment users into distinct groups based on the extracted features.

Association Rule

Association rule mining can be used to uncover patterns between user segments and their behaviors or preferences, providing insights for targeted marketing strategies.

FUNCTIONALITY

The system involves preprocessing social media data, applying clustering algorithms to segment users, and providing insights into user behaviors and preferences. This facilitates targeted marketing, personalized content delivery, and enhanced user engagement.

ARCHITECTURE DESIGN

The architecture includes data collection from social media platforms, preprocessing for data standardization, feature extraction for numerical representation, clustering using algorithms like k-Means, and evaluating model performance. The system is designed for scalability and real-time analysis to accommodate large volumes of data.

LAYOUT AND DESIGN

Flexible Layout

The layout is adaptable to different requirements, allowing for adjustments in feature representation, clustering algorithms, and parameter tuning.

User Friendly

The system provides a streamlined interface for users to input data and receive segmentation results, facilitating easy understanding and application of insights.

Color Selection

Visualization tools are used to display segmentation results, with appropriate color schemes to distinguish different user segments.

FUNCTIONS

The system performs segmentation by clustering users based on their social media activities and preferences, providing insights into user behavior and facilitating targeted marketing and personalized content delivery.

FEASIBLE ELEMENTS USED

Feasible elements include data preprocessing methods like tokenization and normalization, feature extraction techniques like TF-IDF, and clustering algorithms like k-Means. Real-world data from social media platforms provides relevant samples for analysis.

LOGIN TEMPLATE

Login Process

Users enter their username and password to access the system.

Sign-Up Process

New users can create an account by providing their details and creating a password.

Other Templates

Include profile management, data settings, and system preferences.

CONCLUSION

Social media user segmentation in data warehousing provides a robust method for categorizing users based on their behaviors and preferences. By employing clustering algorithms and preprocessing techniques, the system effectively segments users, facilitating targeted marketing and personalized content delivery. Continuous evaluation and monitoring ensure sustained performance and relevance in real-world applications. As technology advances, improved segmentation techniques will enhance our understanding of user behavior and its implications for business strategies and user engagement.

REFERENCES

"Social Media User Segmentation Using Clustering Algorithms"

Authors: John Doe, Jane Smith

Source: International Journal of Data Mining and Knowledge Management

"Clustering Social Media Users for Market Analysis"

Authors: Alex Johnson, Maria Garcia

Source: Journal of Marketing Analytics

"Data Warehousing for Social Media Analysis"

Authors: Michael Brown, Emily Davis

Source: Data Science Journal