Social Media Data Clustering Project

This project involves analyzing and clustering social media post data using various machine learning techniques, primarily focusing on the K-Means clustering algorithm. The dataset used contains information about different types of social media posts, including metrics like reactions, comments, shares, and other user engagement metrics.

Project Overview

The goal of this project is to explore the provided dataset and apply machine learning techniques to cluster similar posts. The process involves data cleaning, exploratory data analysis (EDA), feature engineering, and using clustering methods to group posts with similar characteristics.

Dataset

The dataset used for this project contains **7050 entries** and **16 columns**. Each entry represents a different social media post and provides details such as:

- status_id: Unique identifier for each post
- status_type: The type of post (e.g., video, photo, link, status)
- status_published: The publication date of the post
- num_reactions, num_comments, num_shares: Engagement metrics for each post

Unnecessary columns (Column1, Column2, Column3, Column4) were dropped from the dataset to simplify analysis.

Features

1. **Data Cleaning**: Removed unnecessary columns, checked for missing values, and dropped non-informative features.
2. **Data Transformation**: Categorical variables were encoded into numerical values using LabelEncoder.
3. **Feature Scaling**: The features were scaled using MinMaxScaler.
4. **K-Means Clustering**:
    - A K-Means model was used to cluster the posts into different groups based on the scaled features.
    - The elbow method was used to determine the optimal number of clusters.

Key Dependencies

- Python
- Pandas
- Scikit-learn
- Matplotlib

Installation

To set up this project locally, follow these steps:

1. Clone the repository:

sh

Copy code

git clone https://github.com/yourusername/social-media-clustering.git

2. Install the required packages:

sh

Copy code

pip install -r requirements.txt

Running the Project

You can run the project notebook using Google Colab or Jupyter Notebook. To reproduce the clustering results:

1. Load the dataset into the notebook.
2. Follow the data preprocessing and clustering steps provided in the notebook.
3. Modify parameters as needed to experiment with different models.

Project Results

- A K-Means clustering model with **two clusters** achieved an accuracy of **61%** when matching the predicted cluster labels to the original post type labels.
- Increasing the number of clusters to **four** resulted in a slight decrease in the model's accuracy, suggesting that two clusters provide a better fit for this dataset.