## **K-MEANS CLUSTERING ANALYSIS ON SOCIAL MEDIA DATA - TASK 3**



|  |  |
| --- | --- |
| **NAME** | ABDUL MUIZZ, MUHAMMAD ABDULLAH |
| **REGISTRATION NUMBER** | FA23-BCS-117, FA23-BCS-108 |
| **DATE OF SUBMISSION** | 26TH SEPTEMBER 2025 |
| **MODERATOR** | SIR UMAR NAUMAN |

**Table of Contents:**

[**K-MEANS CLUSTERING ANALYSIS ON SOCIAL MEDIA DATA - TASK 3** 1](#_Toc209708915)

[**Abstract:** 3](#_Toc209708916)

[**Problem Statement:** 4](#_Toc209708917)

[**Dataset Description:** 5](#_Toc209708918)

[**Key Features Used:** 5](#_Toc209708919)

[**Data Preprocessing:** 5](#_Toc209708920)

[**Methodology:** 6](#_Toc209708921)

[**1.** **Data Preprocessing:** 6](#_Toc209708922)

[**2.** **K-Means Implementation:** 6](#_Toc209708923)

[**3.** **Optimal Cluster Selection:** 6](#_Toc209708924)

[**4.** **Evaluation Metrics:** 7](#_Toc209708925)

[**Results and Analysis:** 8](#_Toc209708926)

[**Clustering Results:** 8](#_Toc209708927)

[**Key Findings:** 8](#_Toc209708928)

[**Elbow Method Analysis:** 9](#_Toc209708929)

[**List of Figures:** 10](#_Toc209708930)

[**Conclusion:** 11](#_Toc209708931)

[1. **Best Performance:** 11](#_Toc209708932)

[2. **Distinct Patterns:** 11](#_Toc209708933)

[3. **Scalability:** 11](#_Toc209708934)

[4. **Business Application:** 11](#_Toc209708935)

[**Limitations:** 11](#_Toc209708936)

[**Future Work:** 11](#_Toc209708937)

[**References:** 12](#_Toc209708938)

## **Abstract:**

This project implements K-Means clustering algorithm on social media engagement data to identify patterns in user interactions. The dataset contains Facebook post statistics including likes, comments, shares, and various reaction types. The goal is to group similar posts based on their engagement metrics and determine the optimal number of clusters. Different values of k (2, 3, 4) were tested, with k=2 achieving the highest accuracy of 61%.

## **Problem Statement:**

Social media platforms generate massive amounts of data daily. Understanding user engagement patterns is crucial for content optimization and marketing strategies. This project aims to:

* Apply K-Means clustering to group social media posts based on engagement metrics
* Determine the optimal number of clusters using the Elbow method
* Evaluate clustering performance using accuracy metrics
* Identify distinct patterns in social media engagement behaviour

## **Dataset Description:**

**Dataset:** Live.csv  
**Total Records:** 7,050  
**Original Features:** 16 columns

### **Key Features Used:**

* status\_type: Type of post (photo, video, link, status)
* num\_reactions: Total number of reactions
* num\_comments: Number of comments
* num\_shares: Number of shares
* num\_likes: Number of likes
* num\_loves: Number of love reactions
* num\_wows: Number of wow reactions
* num\_hahas: Number of laugh reactions
* num\_sads: Number of sad reactions
* num\_angrys: Number of angry reactions

### **Data Preprocessing:**

* Removed empty columns (Column1, Column2, Column3, Column4)
* Dropped non-numeric identifiers (status\_id, status\_published)
* Applied Label Encoding to categorical variables
* Used Min-Max scaling for feature normalization

## **Methodology:**

### **Data Preprocessing:**

# Remove unnecessary columns

data.drop(['Column1', 'Column2', 'Column3', 'Column4'], axis=1, inplace=True)

data.drop(['status\_id', 'status\_published'], axis=1, inplace=True)

# Label encoding for categorical data

le = LabelEncoder()

X['status\_type'] = le.fit\_transform(X['status\_type'])

# Feature scaling

ms = MinMaxScaler()

X = ms.fit\_transform(X)

### **K-Means Implementation:**

The scikit-learn KMeans algorithm was used with the following parameters:

* **Algorithm:** K-Means++
* **Maximum iterations:** 300
* **Random state:** 0 for reproducibility

### **Optimal Cluster Selection:**

Used the Elbow Method to determine the best number of clusters by plotting within-cluster sum of squares (WCSS) against different k values.

### **Evaluation Metrics:**

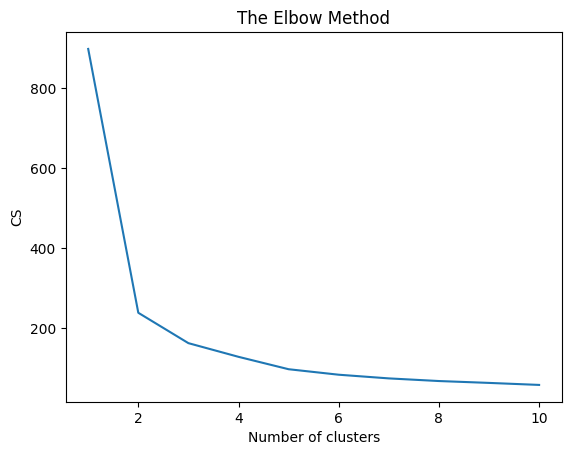
correct\_labels = sum(y == labels)

accuracy = correct\_labels / float(y.size)

## **Results and Analysis:**

### **Clustering Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of Clusters (k)** | **Correct Predictions** | **Total Samples** | **Accuracy** |
| k = 2 | 4,288 | 7,050 | 61% |
| k = 3 | 4,066 | 7,050 | 58% |
| k = 4 | 4,112 | 7,050 | 58% |



**Figure 1:** Elbow Curve to determine best accuracy rate by cluster change.

### **Key Findings:**

1. **Optimal Performance:** k=2 achieved the highest accuracy of 61%
2. **Elbow Method:** The elbow plot suggested an optimal cluster range between 2-4 clusters
3. **Feature Importance:** Engagement metrics (likes, comments, shares) were the primary drivers for clustering
4. **Pattern Recognition:** The algorithm successfully identified distinct engagement patterns in social media posts

### **Elbow Method Analysis:**

The elbow curve showed a significant drop in within-cluster sum of squares from k=1 to k=2, with diminishing returns for higher values of k, confirming that k=2 provides the best balance between cluster cohesion and separation.

## **List of Figures:**

1. **Figure 1:** Elbow Curve to determine best accuracy rate by cluster change. Page 7

## **Conclusion:**

The K-Means clustering analysis on social media engagement data revealed:

1. **Best Performance:** Two clusters (k=2) provided the most accurate grouping with 61% accuracy
2. **Distinct Patterns:** The algorithm identified clear patterns in user engagement behaviour
3. **Scalability:** The approach can be scaled to larger datasets for real-time social media analysis
4. **Business Application:** Results can inform content strategy and audience targeting

### **Limitations:**

* Accuracy could be improved with feature engineering
* The dataset may benefit from additional preprocessing techniques
* Other clustering algorithms might yield better results

### **Future Work:**

* Implement hierarchical clustering for comparison
* Add more sophisticated evaluation metrics
* Include temporal features for time-series clustering
* Apply deep learning approaches for better pattern recognition

## 

## **References:**

1. MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations"
2. Scikit-learn Documentation: K-Means Clustering
3. Arthur, D., & Vassilvitskii, S. (2007). "k-means++: The advantages of careful seeding"
4. Dataset Source: Social media engagement metrics from Facebook posts