

**National University of Computer and Emerging Sciences**



<Clustering Twitter User Activity Using K-Means and Network Visualization>

**Group Members**

<Bilal Ahmed>…………<22K-4525> | Group Leader

<Ashar Zamir>…………<22K-4241>

<Anas Saleem>…………<22K-0500>

**Project Objective:**

The primary objective of this project is to analyze Twitter user activity through clustering techniques, specifically K-Means, to identify different communities of users based on their engagement metrics (retweets, likes, and tweets). The project also involves visualizing these communities using NetworkX to understand the relationships between users based on their activity patterns. Precision, recall, F1-score, and silhouette score are used to evaluate the clustering performance.

**Dataset Overview:**

The dataset used for this project is a collection of Twitter user activity, including metrics like the number of retweets, likes, and total tweets for each user. The dataset is processed to group activity by the username, and a sample of 1000 users is selected for the analysis. The focus is on understanding user behavior patterns by using these engagement metrics.

**Data Processing Steps:**

1. **Data Aggregation:** The data is grouped by the 'Username' to calculate total retweets, total likes, and total tweets per user.
2. **Sampling:** A random sample of 1000 users is selected for faster processing and testing.
3. **Data Splitting:** The dataset is split into training and testing sets. The training set consists of 500 users, and the test set consists of 200 users.
4. **Feature Normalization:** The features (total retweets, total likes, and total tweets) are normalized using StandardScaler to ensure that all features contribute equally to the clustering process.
5. **K-Means Clustering:** The K-Means algorithm is applied to cluster users into two communities based on their activity. The number of clusters (k) is set to 2 for this analysis, but this can be adjusted to explore other configurations.
6. **Evaluation:** The clustering quality is evaluated using the silhouette score, which provides an indication of how well the data is clustered. Precision, recall, and F1-score are calculated to assess the clustering performance in comparison to the test set.

**Visualization:**

NetworkX is used to visualize the clusters. A graph is constructed where each node represents a user, and edges are created based on the similarity of their activity. The nodes are colored based on their assigned community, providing a visual representation of user clusters.

**Clustering Results:**

* **Silhouette Score:** The silhouette score for the test set is calculated to evaluate how well the K-Means algorithm has clustered the users. A higher silhouette score indicates better-defined clusters.
* **Precision, Recall, and F1-Score:** These metrics are used to evaluate the clustering results. Since the dataset does not have predefined labels, these scores compare the predicted clusters with the actual cluster assignments in the test set.
  + **Precision:** The precision score measures the accuracy of the predicted clusters.
  + **Recall:** The recall score indicates how well the model captures all the relevant instances.
  + **F1-Score:** The F1-score provides a balanced measure of precision and recall

.

**Visualization of Communities:**

The graph visualization shows the clustering of users into distinct communities. The force-directed layout positions nodes in a way that groups similar nodes together, and the nodes are colored based on their community label. This helps to visualize the relationships between users based on their activity.

**Conclusion:**

The project successfully demonstrates how K-Means clustering can be applied to Twitter user data to identify different user communities based on engagement metrics. The results are evaluated using various metrics to understand the clustering quality, and the visualization offers a clear depiction of the user communities.