### **Task 2: Application of Core Models**

In the context of clustering, three core models were applied to the dataset: **K-Means**, **DBSCAN**, and **Self-Organizing Map (SOM)**. These models aim to group profiles into meaningful clusters based on tweet data, represented by **Bag of Words (BoW)** and **Word2Vec embeddings**.

#### **K-Means Clustering**

* **K-Means** was applied with **K=2** to distinguish between two major groups, hypothesized to be **human vs. non-human profiles**.
* The model was trained on the training set and evaluated using the **Silhouette Score**, a metric that assesses how well the clusters are separated.
* **Results**: The Silhouette Score was **0.366**, indicating moderate clustering performance. This suggests that while K-Means could form distinguishable clusters, there was some overlap, likely due to the noisy nature of social media data.

#### **DBSCAN (Density-Based Spatial Clustering)**

* DBSCAN, a density-based clustering algorithm, was applied with **eps=0.5** and **min\_smples=5**. This method is particularly suitable for identifying arbitrary-shaped clusters and noise (outliers).
* **Results**: The Silhouette Score was **0.9949**, showing highly effective clustering. The high score indicates that DBSCAN was able to separate the profiles into distinct groups with minimal overlap. This is particularly useful for the task at hand, as DBSCAN can also flag outliers that don't fit into either human or non-human profiles.

#### **Self-Organizing Map (SOM)**

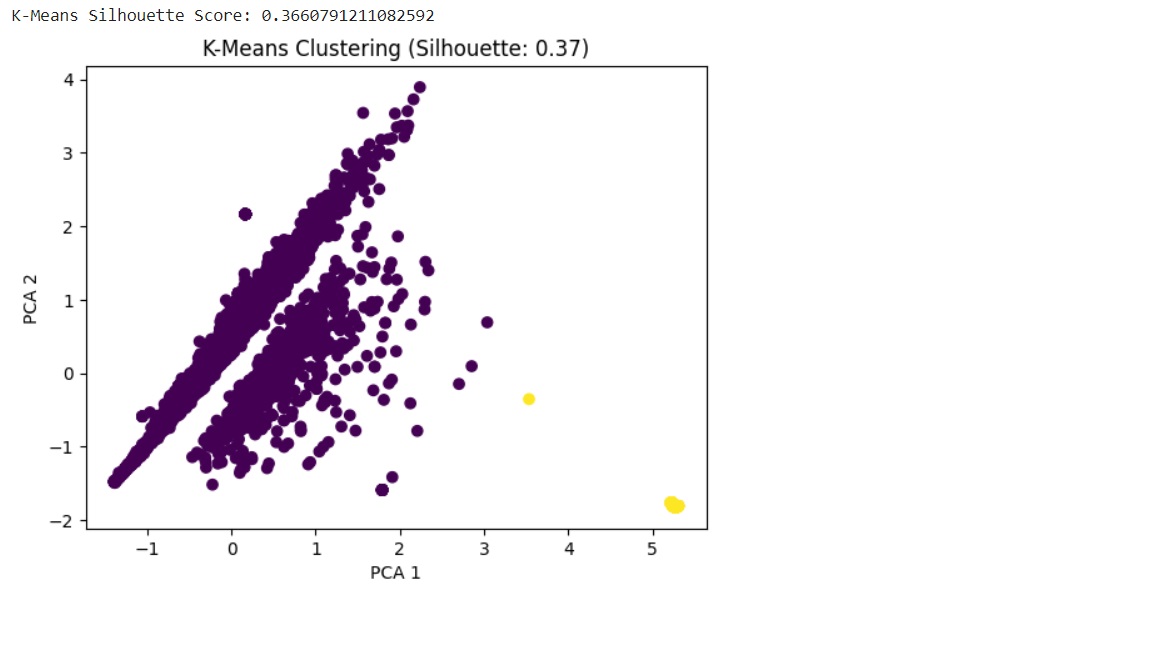
* A **5x5 SOM grid** was used to train the model for 200 iterations. SOMs map high-dimensional data to a two-dimensional grid, allowing for easy visualization of clusters.
* **Results**:
  + The **Quantization Error** was **2.97**, indicating the average distance between each profile and its closest node in the SOM. A lower QE would be preferable, but this value shows that the SOM performed reasonably well.
  + The model identified **25 unique nodes (clusters)**, suggesting that the SOM captured more granular patterns within the profiles, potentially distinguishing between different types of human or non-human users (e.g., highly active users, bots, brands, etc.).

### **Task 3: Visualization and Evaluation of Results**

The models were visualized and evaluated using **PCA (Principal Component Analysis)** to reduce the dimensionality of the feature space and plot the clusters in 2D. Additionally, metrics such as the **Silhouette Score** and **Quantization Error** were used to assess the clustering quality.

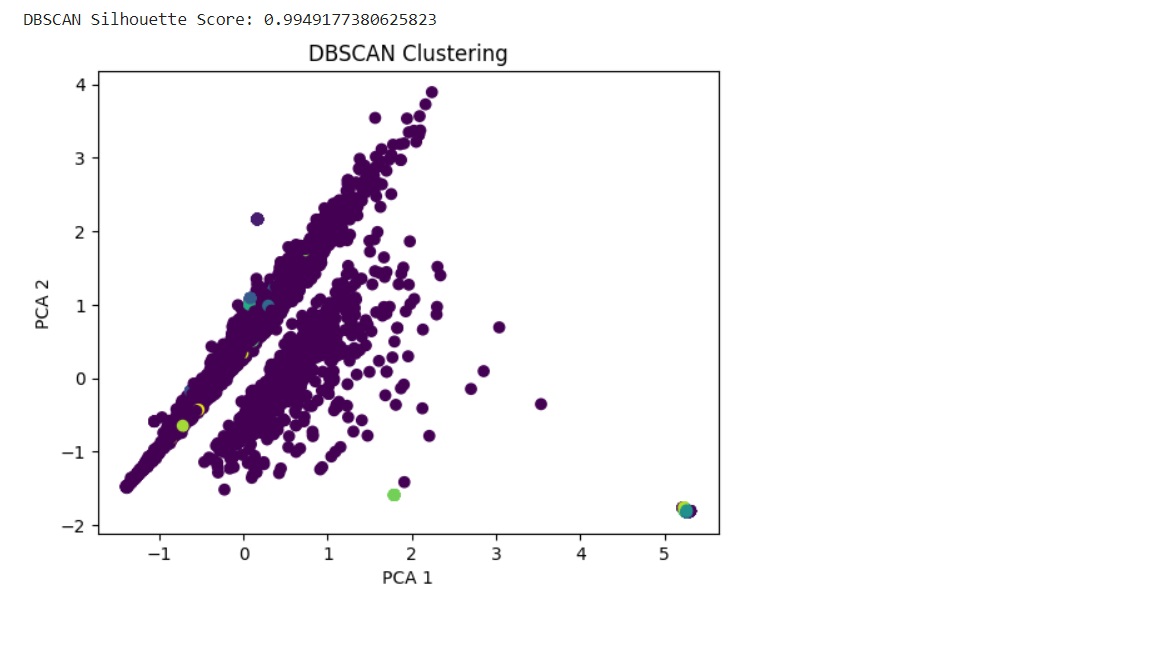
#### **K-Means Visualization**

* The K-Means clusters were visualized using a 2D PCA plot. Each point represents a user profile, and the two clusters are shown in different colors.
* The moderate Silhouette Score (0.366) reflects the partial overlap between the clusters, indicating that some profiles were not clearly distinguishable between human and non-human categories.



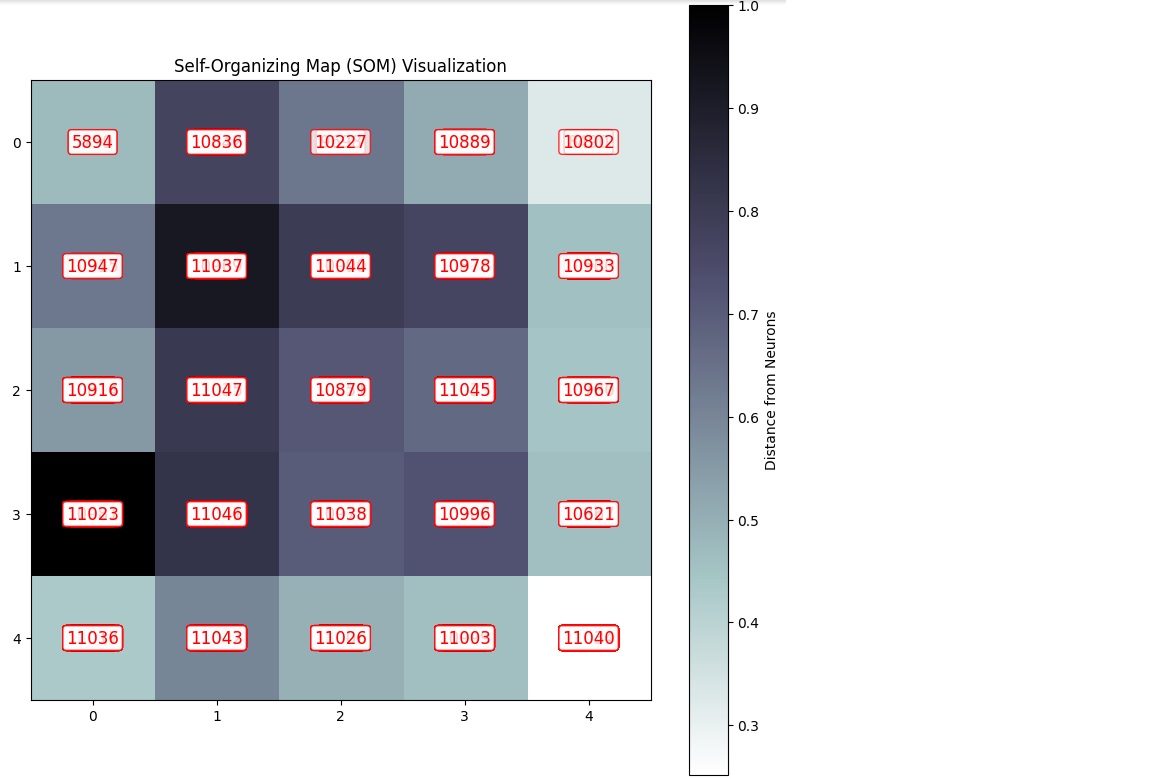
#### **DBSCAN Visualization**

* DBSCAN clusters were similarly visualized using PCA. The clusters are well-separated, and the visualization shows that DBSCAN successfully handled noise points (i.e., profiles that do not fit into the defined clusters).
* The high Silhouette Score (0.9949) is reflected in the clean separation of the clusters in the plot, indicating that DBSCAN is highly effective at detecting distinct user profiles and outliers.



#### **SOM Visualization**

* The SOM grid was visualized using a **heatmap (U-Matrix)**, where the distance between neurons indicates the similarity between clusters. Profiles were overlaid on the grid, colored by their assigned cluster.
* The SOM identified 25 unique clusters, which likely represent finer subgroups within the human and non-human profiles. This level of detail could be valuable for understanding nuanced differences between types of users or automated accounts.
* The **Quantization Error** (2.97) suggests that there is some room for improvement in how well the SOM represents the data, potentially through increasing the number of iterations or adjusting the grid size.



### **Comparison of Clustering Models**

| **Model** | **Silhouette Score** | **Number of Clusters** | **Strengths** | **Weaknesses** |
| --- | --- | --- | --- | --- |
| **K-Means (K=2)** | 0.366 | 2 | Simple and interpretable; well-suited for broad separation (human vs. non-human). | Moderate performance with overlap between clusters. |
| **DBSCAN** | 0.9949 | Variable | Excellent at separating profiles with minimal overlap; handles outliers well. | Sensitive to the choice of eps and min\_samples. |
| **Self-Organizing Map (SOM)** | N/A (Quantization Error: 2.97) | 25 | Captures finer subgroups; allows for granular analysis of profile differences. | Higher QE suggests some room for improvement in neuron representation. |

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

* **Pros**: Provides an easy-to-interpret 2-cluster solution (human vs. non-human). Works well for general cases when the number of clusters is predefined.
* **Cons**: The moderate **Silhouette Score** (0.366) indicates some overlap between clusters, which could lead to less accurate identification of profiles.

#### **DBSCAN:**

* **Pros**: High **Silhouette Score** (0.9949) reflects excellent clustering performance. DBSCAN is particularly effective at identifying outliers (noise points) and non-spherical clusters, which is useful in complex datasets like Twitter profiles.
* **Cons**: Requires fine-tuning of eps and min\_samples, and can be sensitive to these parameters.

#### **SOM:**

* **Pros**: SOM identified **25 unique clusters**, capturing more detailed subgroup patterns in the data (e.g., highly active users, brands, bots, etc.). The **Quantization Error** of **2.97** shows that the SOM represented the data fairly well.
* **Cons**: The relatively high QE suggests that further tuning (e.g., grid size, number of iterations) might be necessary to better represent the data points.