Report on Text Data Analysis and DBSCAN Clustering

#### **Overview**

This report details the analysis of textual data using advanced machine learning techniques. The aim was to classify and cluster messages based on their Jaccard similarity scores and to present the results through dimensionality reduction and visualization.

#### **Data Preparation**

Two datasets were involved in the analysis:

1. **LLM Binary Complete Dataset:** Includes binary features with Jaccard similarity scores.
2. **Ushahidi Message Dataset:** Contains text descriptions of messages.

The first step involved loading these datasets and preprocessing the Jaccard similarity scores to create binary classifications of zero and non-zero similarity scores.

#### **Jaccard Score Classification**

Jaccard similarity scores were categorized into two groups:

* **Zero Jaccard Scores:** Representing no similarity.
* **Non-zero Jaccard Scores:** Representing some level of similarity.

This categorization was essential for encoding and clustering the messages.

#### **Message Encoding**

Messages from the Ushahidi dataset were transformed into numerical vectors using the T5 Encoder Model. This model encodes textual data into a format suitable for clustering, resulting in vectors of fixed dimensions.

#### **Clustering Analysis**

The encoded messages were clustered based on the Jaccard score categories:

* **Zero Jaccard Scores:** These messages were grouped into 7 distinct clusters.
* **Non-zero Jaccard Scores:** These messages were grouped into 3 distinct clusters.

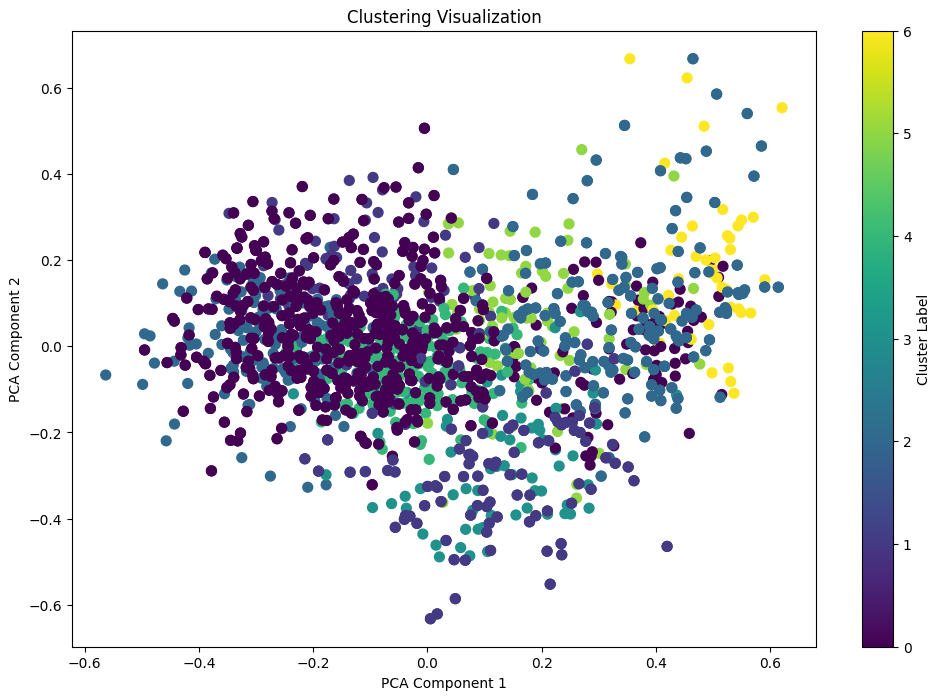
These clusters reflect different thematic groupings within the data.

#### **Dimensionality Reduction and Visualization**

To facilitate the visualization of clustering results, Principal Component Analysis (PCA) was used to reduce the dimensionality of the encoded message vectors to two dimensions. A scatter plot was created to represent each message in this 2D space, color-coded by cluster.

This visualization helped in understanding the distribution and patterns of the clusters.

#### **Results**



* **Shape of llm\_df:** (1472, 76)

This indicates that the llm\_df dataframe has 1472 rows and 76 columns after processing Jaccard scores.

* **Shape of encoded\_messages:** (2020, 512)

The encoded message vectors have 2020 rows and 512 dimensions.

* **Cluster Counts:**
  + **Zero Jaccard Scores:** 7 high-level concept clusters.
  + **Non-zero Jaccard Scores:** 3 high-level concept clusters.

#### **Conclusion**

The analysis successfully categorized and clustered textual data based on Jaccard similarity scores. Through encoding, clustering, and dimensionality reduction, clear insights into the data structure and cluster distribution were obtained. These results provide a foundation for further exploration and application in natural language processing tasks.