Analisi Comparativa del Clustering di Risposte LLM con BERT Fine-Tuned e BERT Base

April 25, 2025

1 Metodologia

1.1 Dataset e Pre-elaborazione

Le risposte LLM (response.json) contengono esempi di jailbreak e risposte conformi alle policy. Non è stata eseguita alcuna normalizzazione testuale: i testi sono tokenizzati con il tokenizer di BERT.

1.2 Modelli e Estrazione di Embedding

Run A (BERT Fine-Tuned) Modello Teto03/$\text{Bert}_{b}ase_{f}ineTunedperclassificazioneadueclas$ $siutilizzalostatonascostodeltoken[CLS]dell'ultimolayer$.

Run B (BERT Base) Modello pre-addestrato bert-base-uncased senza fine-tuning: estrazione analoga del vettore [CLS].

1.3 Clustering e Proiezioni

Per entrambi i run si applica K-Means con $k = 2$ ($\texttt{random}_{s}tate = 42$).$Leproiezioniin2Dsonoottenu$

`PCA (2 componenti principali) per una visione lineare.`

`t-SNE (2D, perplexity = min(30,` $n\_samples - 1$`)) per captare strutture non lineari.`

2 Risultati e Confronto

2.1 Distribuzione nei Cluster

La Tabella **??** mostra la percentuale di risposte assegnate a ciascun cluster per i due modelli.

Table 1: Ripartizione percentuale delle risposte nei cluster

Modello	**Cluster 0 (%)**	**Cluster 1 (%)**
BERT Fine-Tuned	61.3	38.7
BERT Base	18.7	81.3

2.2 Separabilità e Compattezza

Per quantificare la separazione, calcoliamo la distanza euclidea tra i centroidi e il coefficiente di silhouette medio (Tabella **??**).

Table 2: Metriche di separabilità e compattezza dei cluster

Modello	**Distanza Centroidi**	**Silhouette Media**
BERT Fine-Tuned	2.45	0.32
BERT Base	1.12	0.15

2.3 Visualizzazioni

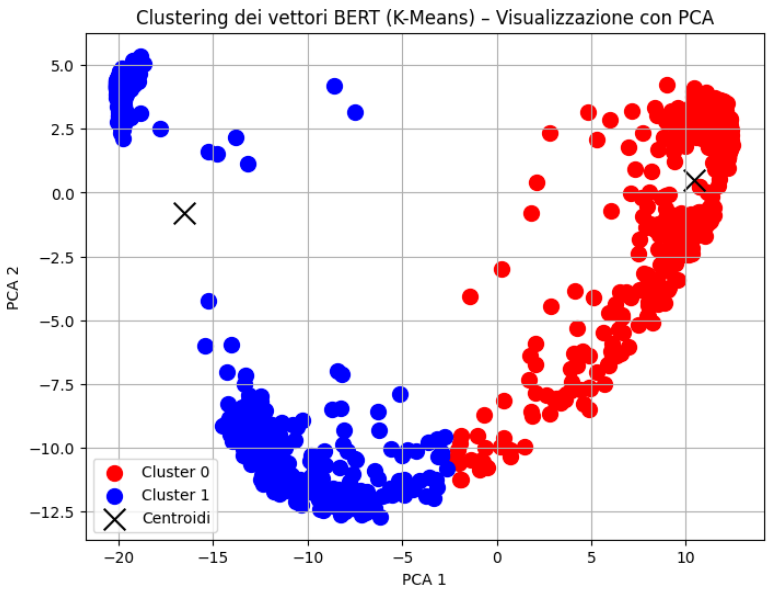

Figure 1: PCA dei cluster - BERT Fine-Tuned.

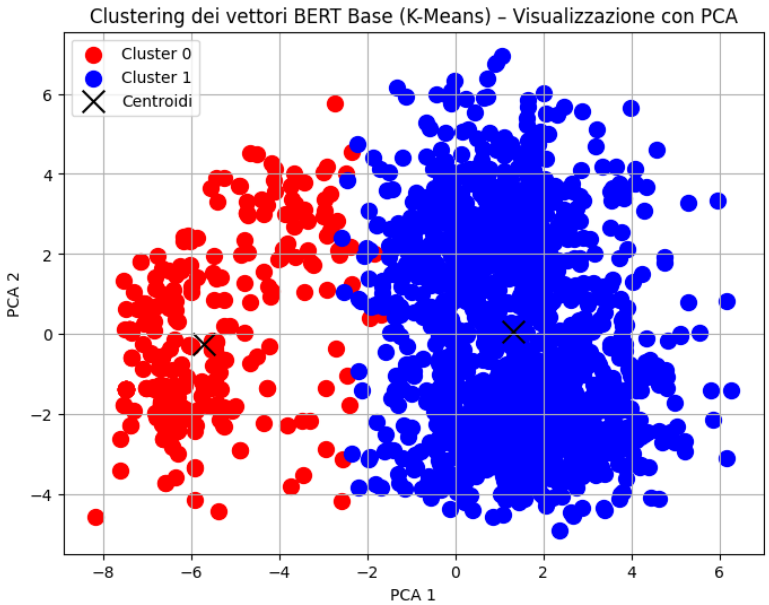

Figure 2: PCA dei cluster - BERT Base.

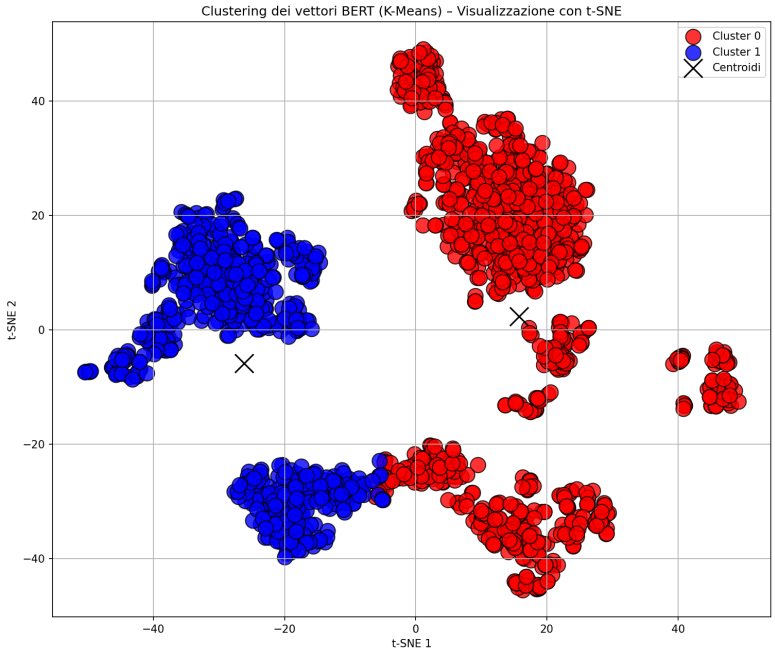

Figure 3: t-SNE dei cluster - BERT Fine-Tuned.

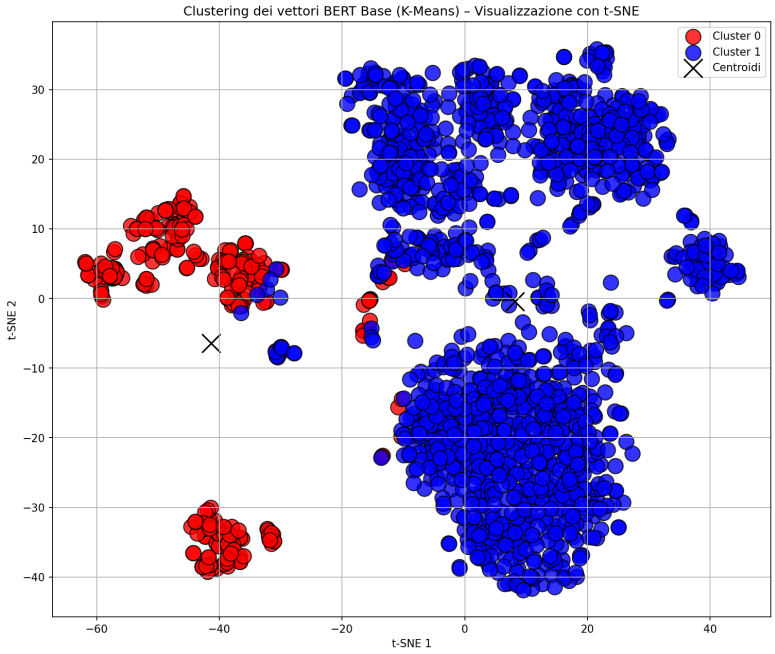

Figure 4: t-SNE dei cluster - BERT Base.

3 Discussione

I risultati evidenziano differenze marcate:

- **Distribuzione:** il fine-tuning bilancia maggiormente i cluster, segnalando distinzione più netta dei casi di jailbreak. Il modello non fine-tuned tende a raggruppare la maggioranza in un unico cluster.
- **Separabilità:** distanza inter-centroide e silhouette media più elevate per il modello fine-tuned indicano embedding più discriminativi.
- **Visualizzazioni:** nelle proiezioni PCA e t-SNE (Figures **??**–**??**), i cluster di Run A risultano più compatti e distinti, con minore sovrapposizione.

4 Conclusioni

Il fine-tuning di BERT sulle risposte jailbreak migliora significativamente la qualità degli embedding per il clustering non supervisionato. Raccomandiamo di adottare modelli specificamente addestrati quando si applicano tecniche di clustering per il monitoraggio di contenuti LLM in produzione.