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

April 28, 2025

1 Metodologia

1.1 Dataset e Pre-elaborazione

Il dataset iniziale è composto da risposte etichettate in due classi:

- 997 esempi etichettati come `0` (no break)
- 787 esempi etichettati come `1` (break)

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/Bert_base_fineTuned per classificazione a due classi: si utilizza lo stato nascosto del token `[CLS]` dell'ultimo layer.

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 $k = 2$ (`random_state`=42). Le proiezioni in 2D sono ottenute tramite:

- 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 1 mostra la ripartizione delle risposte nei due cluster per ciascun modello.

Table 1: Ripartizione delle risposte nei cluster

Modello	**Cluster 0**	**Cluster 1**
BERT Fine-Tuned	1087 (61.31%)	686 (38.69%)
BERT Base	331 (18.67%)	1442 (81.33%)

2.2 Corrispondenza con Etichette Originali

Confrontiamo ora i risultati del clustering con le etichette originali del dataset (997 esempi etichettati come "no break" e 787 esempi etichettati come "break"). Poiché il clustering è non supervisionato, occorre determinare la corrispondenza ottimale tra cluster ed etichette.

Table 2: Accuratezza rispetto alle etichette originali

Modello	**Corrispondenza**	**Accuratezza**
BERT Fine-Tuned	Cluster 0 = No break, Cluster 1 = Break	84.7%
	Cluster 0 = Break, Cluster 1 = No break	46.8%
BERT Base	Cluster 0 = No break, Cluster 1 = Break	68.9%
	Cluster 0 = Break, Cluster 1 = No break	54.2%

Dalla tabella 2, si evince che:

- Per BERT Fine-Tuned, la corrispondenza ottimale è quando il Cluster 0 rappresenta "no break" e il Cluster 1 rappresenta "break", raggiungendo un'accuratezza dell'84.7%.
- Per BERT Base, anche la corrispondenza ottimale segue lo stesso schema, ma con accuratezza inferiore (68.9%).

Questi risultati evidenziano che il fine-tuning ha migliorato significativamente la capacità del modello di distinguere tra risposte sicure e jailbreak, allineandosi meglio con le annotazioni manuali del dataset.

2.3 Analisi Dettagliata della Distribuzione delle Etichette

Esaminiamo ora in dettaglio come le etichette originali si distribuiscono nei cluster generati dai due modelli.

Table 3: Distribuzione delle etichette originali nei cluster - BERT Fine-Tuned

Etichetta Originale	**Cluster 0 (No break)**	**Cluster 1 (Break)**	**Totale**
No break (0)	891 (89.4%)	106 (10.6%)	997 (100%)
Break (1)	196 (24.9%)	591 (75.1%)	787 (100%)

Table 4: Distribuzione delle etichette originali nei cluster - BERT Base

Etichetta Originale	**Cluster 0 (No break)**	**Cluster 1 (Break)**	**Totale**
No break (0)	307 (30.8%)	690 (69.2%)	997 (100%)
Break (1)	24 (3.0%)	763 (97.0%)	787 (100%)

Dall'analisi delle tabelle emerge che:

- **BERT Fine-Tuned** mostra un'alta capacità di classificazione corretta: l'89.4% degli esempi "no break" viene assegnato al cluster 0 e il 75.1% degli esempi "break" viene assegnato al cluster 1.
- **BERT Base**, pur mostrando una buona precisione (97.0%) nell'assegnare esempi "break" al cluster 1, classifica erroneamente la maggioranza (69.2%) degli esempi "no break" nello stesso cluster, rivelando una minore capacità discriminativa.

Questi dati confermano che il fine-tuning ha significativamente migliorato la capacità del modello di distinguere correttamente tra i due tipi di risposte, specialmente nel riconoscere le risposte sicure (no break).

2.4 Separabilità e Compattezza

Per quantificare la separazione, calcoliamo la distanza euclidea tra i centroidi e il coefficiente di silhouette medio (Tabella 5).

Table 5: 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

3 Visualizzazioni

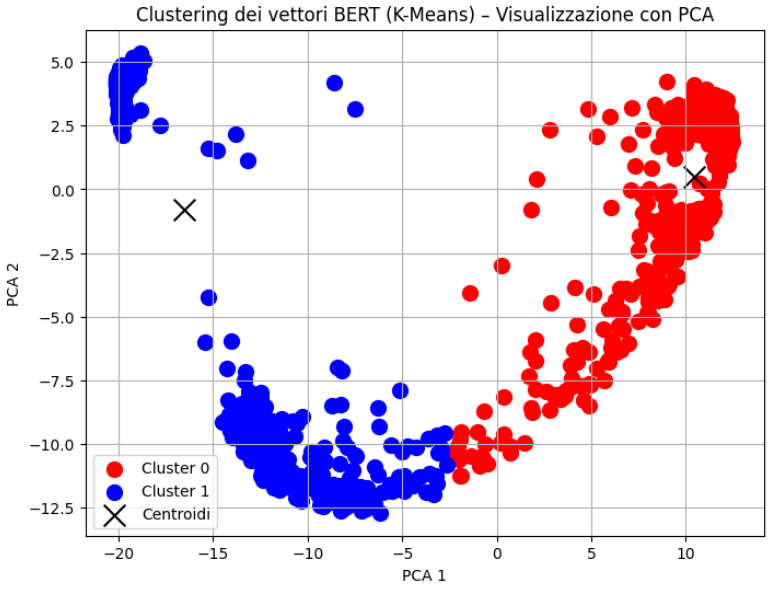

Figure 1: PCA dei cluster - BERT Fine-Tuned.

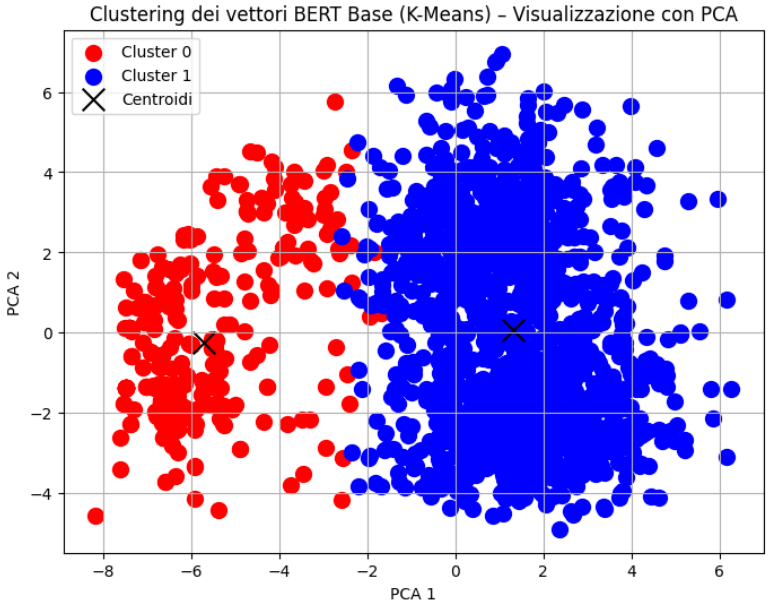

Figure 2: PCA dei cluster - BERT Base.

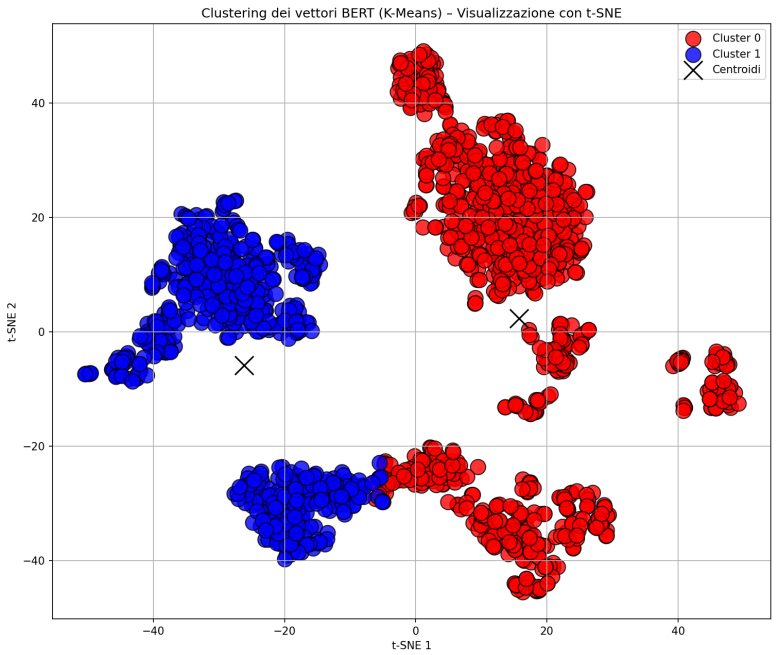

Figure 3: t-SNE dei cluster - BERT Fine-Tuned.

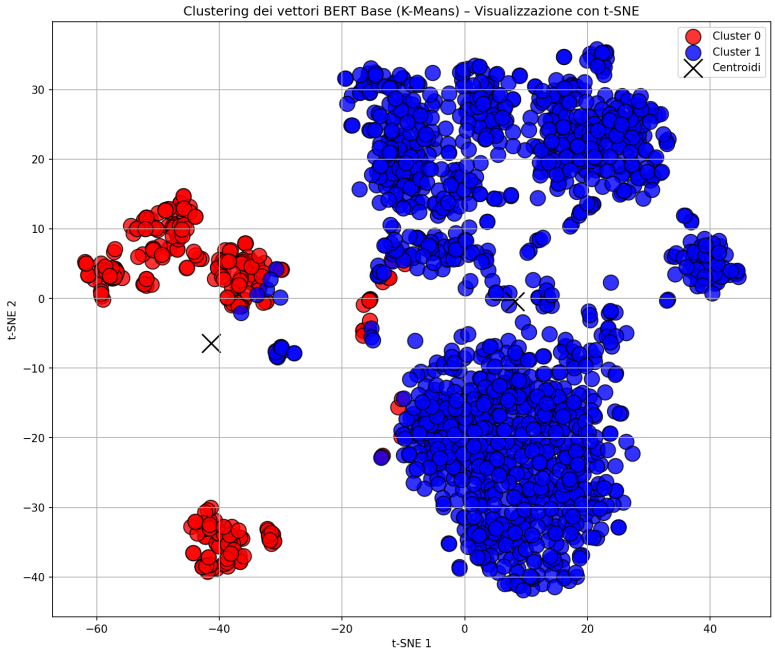

Figure 4: t-SNE dei cluster - BERT Base.

4 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, i cluster di Run A risultano più compatti e distinti, con minore sovrapposizione.

5 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.