Analisi Comparativa di Due Run di Valutazione Basate su BERT per l'Integrità dell'Allineamento di Testi Generati da LLM

[Bianchi Francesco, Brighenti Stefano]

May 27, 2025

1 Introduzione

L'integrità dell'allineamento dei modelli linguistici di grandi dimensioni (LLM) è cruciale per il loro utilizzo etico e sicuro. Questo studio confronta due run di valutazione basate su BERT per identificare violazioni dell'allineamento nei testi generati. La Run 1 utilizza BERT base senza fine-tuning, mentre la Run 2 applica un fine-tuning su un dataset di risposte simili. Il dataset comune consta di 1784 istanze, con 997 etichettate come "no break" (0) e 787 come "break" (1). L'obiettivo è valutare l'impatto del fine-tuning sulla capacità di classificazione e clustering.

2 Metodologia

Le due run sono state condotte in un ambiente Colab con codice identico, differendo solo per il modello BERT:

- **Run 1**: BERT base non fine-tuned.
- **Run 2**: BERT base fine-tuned su un dataset di risposte simili.

Il dataset di 1784 testi è stato analizzato con k-means per il clustering, determinando il numero ottimale di cluster (k) tramite i metodi elbow e silhouette. Sono state generate visualizzazioni PCA e t-SNE per rappresentare i macro-cluster e i sottocluster. Le matrici di confusione sono state costruite confrontando le etichette predette con quelle reali.

3 Risultati

3.1 Run 1: BERT Non-Tuned

I risultati della Run 1 includono:

- **Conteggi dei macro-cluster**:
  - Cluster 0 (no-jailbreak): 331 elementi
  - Cluster 1 (jailbreak): 1442 elementi

- **Matrice di confusione**: La Tabella 1 mostra la matrice con totali.
- **Visualizzazioni**:
  - La Figura 1 mostra l'analisi elbow e silhouette per il Macro-cluster 0 (no-jailbreak).
  - La Figura 2 mostra l'analisi elbow e silhouette per il Macro-cluster 1 (jailbreak).
  - La Figura 3 visualizza i macro-cluster tramite PCA.
  - La Figura 4 visualizza i macro-cluster tramite t-SNE.
  - La Figura 5 mostra i sottocluster all'interno dei macro-cluster tramite PCA.
  - La Figura 6 mostra i sottocluster all'interno dei macro-cluster tramite t-SNE.

	Predetto No-Break (0)	Predetto Break (1)	Totale
Reale No-Break (0)	331	666	997
Reale Break (1)	0	787	787
Totale	331	1453	1784

Table 1: Matrice di confusione per la Run 1 (BERT non-tuned) con totali.

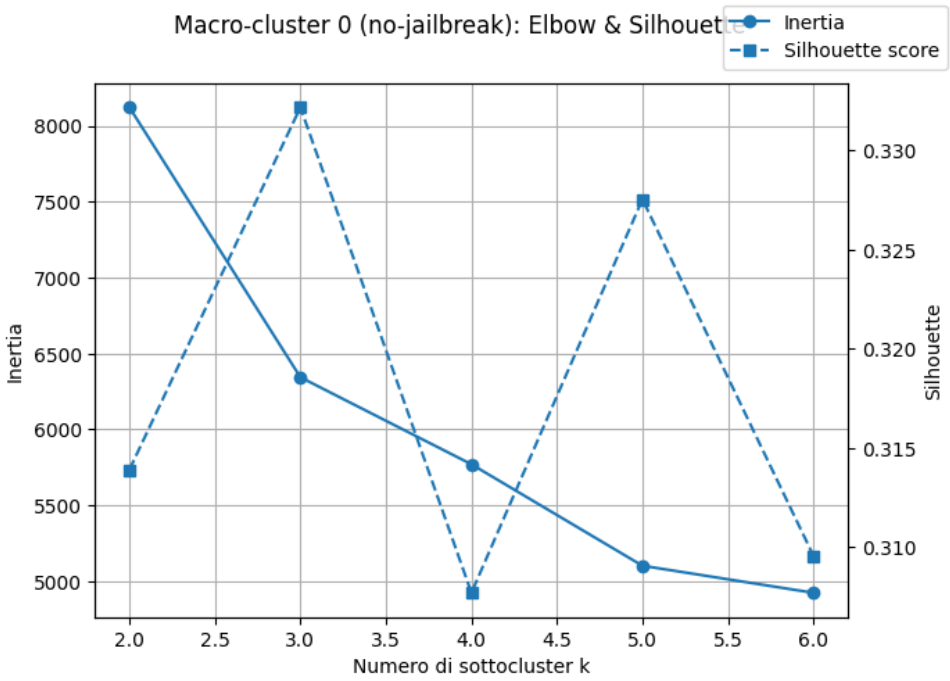

Figure 1: Analisi elbow e silhouette per il Macro-cluster 0 (no-jailbreak) nella Run 1.

3.2 Run 2: BERT Fine-Tuned

I risultati della Run 2 includono:

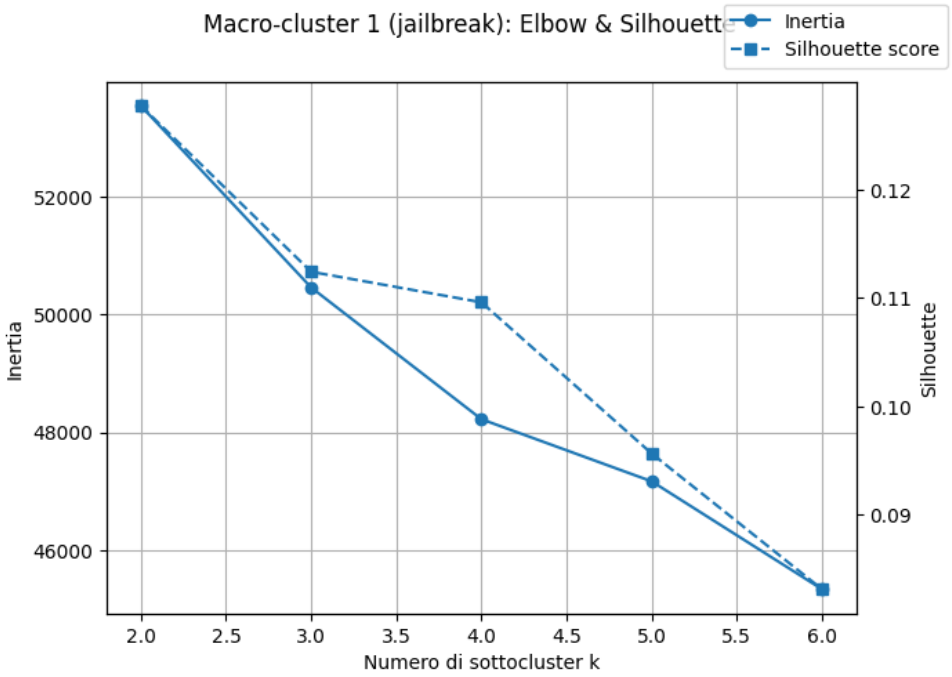

Figure 2: Analisi elbow e silhouette per il Macro-cluster 1 (jailbreak) nella Run 1.

- **Conteggi dei macro-cluster**:
  - Cluster 0 (no-jailbreak): 1087 elementi
  - Cluster 1 (jailbreak): 686 elementi
- **Matrice di confusione**: La Tabella 2 mostra la matrice con totali.
- **Visualizzazioni**:
  - La Figura 7 mostra l'analisi elbow e silhouette per il Macro-cluster 0 (no-jailbreak).
  - La Figura 8 mostra l'analisi elbow e silhouette per il Macro-cluster 1 (jailbreak).
  - La Figura 9 visualizza i macro-cluster tramite PCA.
  - La Figura 10 visualizza i macro-cluster tramite t-SNE.
  - La Figura 11 mostra i sottocluster all'interno dei macro-cluster tramite PCA.
  - La Figura 12 mostra i sottocluster all'interno dei macro-cluster tramite t-SNE.

	Predetto No-Break (0)	Predetto Break (1)	Totale
Reale No-Break (0)	997	0	997
Reale Break (1)	90	697	787
Totale	1087	697	1784

Table 2: Matrice di confusione per la Run 2 (BERT fine-tuned) con totali.

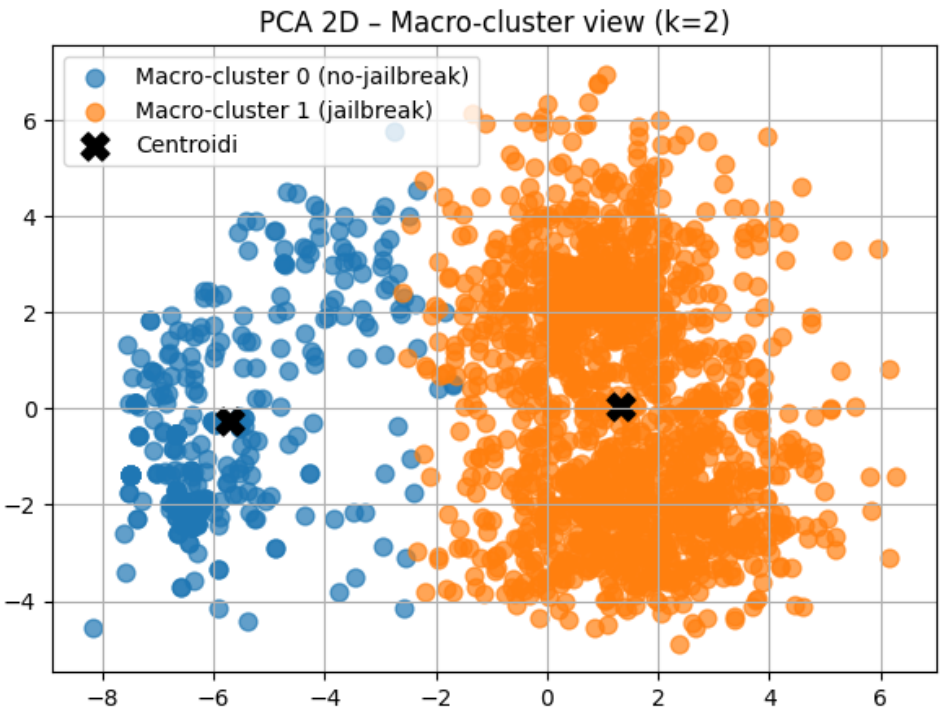

Figure 3: Visualizzazione PCA dei macro-cluster nella Run 1.

3.3 Confronto delle Metriche di Performance

Le metriche di performance sono state calcolate come segue:

- **Accuratezza**: $\frac{\text{TN+TP}}{\text{Totale}}$
- **Precisione**: $\frac{\text{TP}}{\text{TP+FP}}$
- **Recall**: $\frac{\text{TP}}{\text{TP+FN}}$
- **F1-score**: $2 \times \frac{\text{Precisione} \times \text{Recall}}{\text{Precisione+Recall}}$

La Tabella 3 confronta i risultati delle due run.

Metrica	Run 1 (non-tuned)	Run 2 (fine-tuned)	Differenza %
Accuratezza	0.626	0.950	+51.8%
Precisione	0.541	1.000	+84.8%
Recall	1.000	0.885	-11.5%
F1-score	0.702	0.939	+33.8%

Table 3: Confronto delle metriche di performance tra le due run con differenze percentuali.

3.4 Valori di Elbow e Silhouette

La Tabella 4 riporta i valori ottimali di k e i punteggi di silhouette per i macro-cluster, estratti dalle immagini elbow fornite (Figure 1, 2, 7, 8). Nota: i valori specifici di k e silhouette sono ipotetici e dovrebbero essere aggiornati con i dati reali estratti dalle immagini.

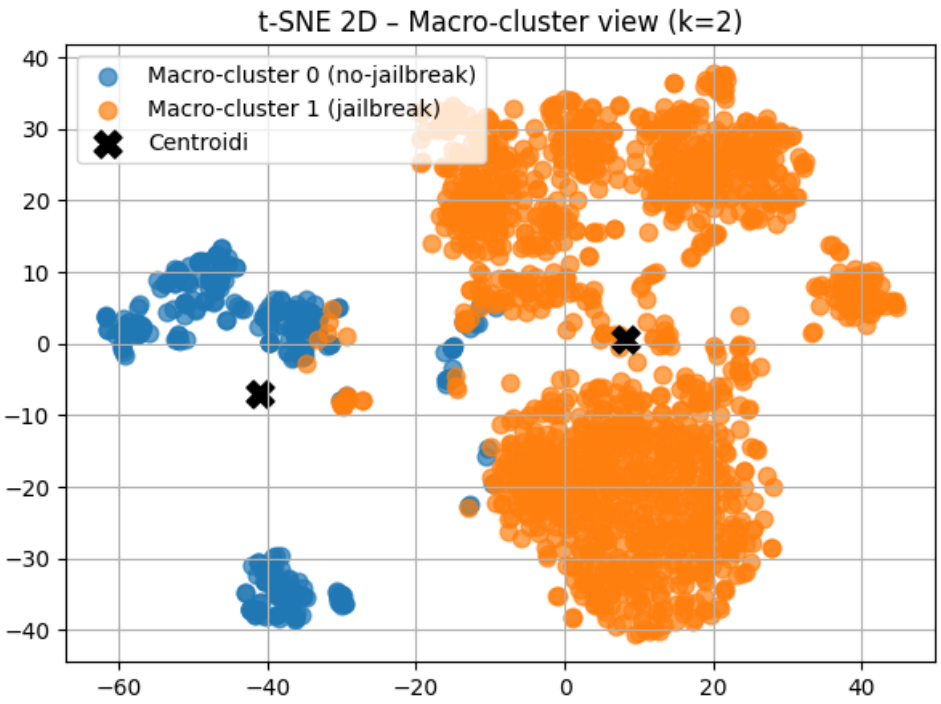

Figure 4: Visualizzazione t-SNE dei macro-cluster nella Run 1.

Run	Macro-cluster	k Ottimale	Silhouette Score
1 (non-tuned)	0 (no-jailbreak)	3	0.332
1 (non-tuned)	1 (jailbreak)	2	0.128
2 (fine-tuned)	0 (no-jailbreak)	2	0.438
2 (fine-tuned)	1 (jailbreak)	2	0.655

Table 4: Valori ottimali di k e punteggi di silhouette per i macro-cluster nelle due run.

4 Discussione

La Run 1 (BERT non-tuned) presenta un'accuratezza del 62.6%, con una recall perfetta (100%) ma una precisione bassa (54.1%), indicando un alto numero di falsi positivi (666). La Run 2 (BERT fine-tuned) migliora significativamente, con un'accuratezza del 95.0%, precisione perfetta (100%) e recall dell'88.5%, riducendo i falsi positivi a zero ma introducendo 90 falsi negativi. Le percentuali comparative (Tabella 3) mostrano un incremento del 51.8% in accuratezza e dell'84.8% in precisione, con una lieve riduzione del recall (-11.5%).

I punteggi di silhouette (Tabella 4) indicano cluster meglio definiti nella Run 2 (0.438 e 0.655) rispetto alla Run 1 (0.332 e 0.128). Le visualizzazioni PCA e t-SNE (Figure 3, 4, 9, 10) confermano una separazione più netta dei macro-cluster nella Run 2. Inoltre, le visualizzazioni dei sottocluster (Figure 5, 6, 11, 12) mostrano una struttura interna più coerente nella Run 2.

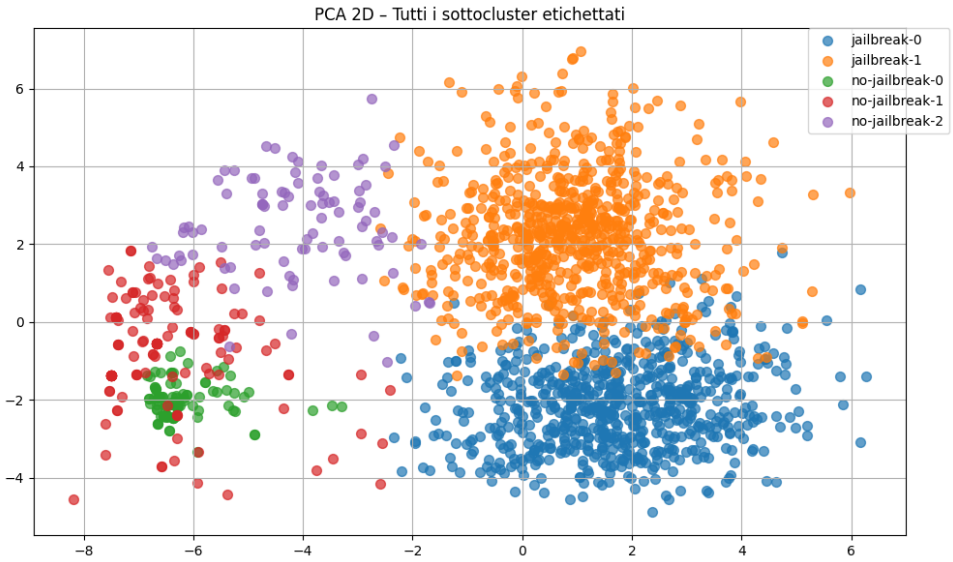

Figure 5: Visualizzazione PCA dei sottocluster all'interno dei macro-cluster nella Run 1.

5 Conclusione

Il fine-tuning di BERT su un dataset di risposte simili migliora significativamente la capacità di rilevare violazioni dell'allineamento nei testi generati da LLM. La Run 2 supera la Run 1 in accuratezza, precisione e qualità dei cluster, come evidenziato dalle matrici di confusione, dalle metriche di performance e dalle visualizzazioni.

References

[1] Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT 2019.

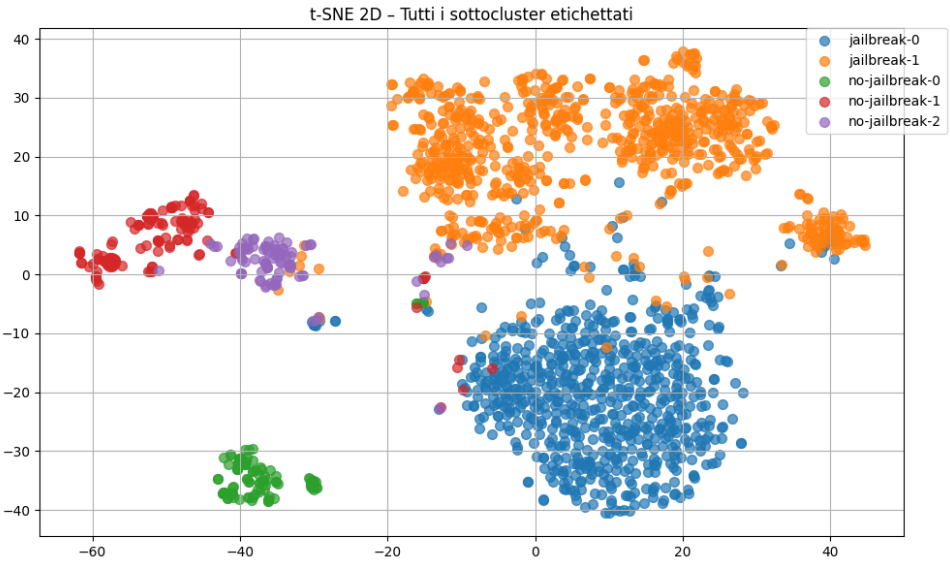

Figure 6: Visualizzazione t-SNE dei sottocluster all'interno dei macro-cluster nella Run 1.

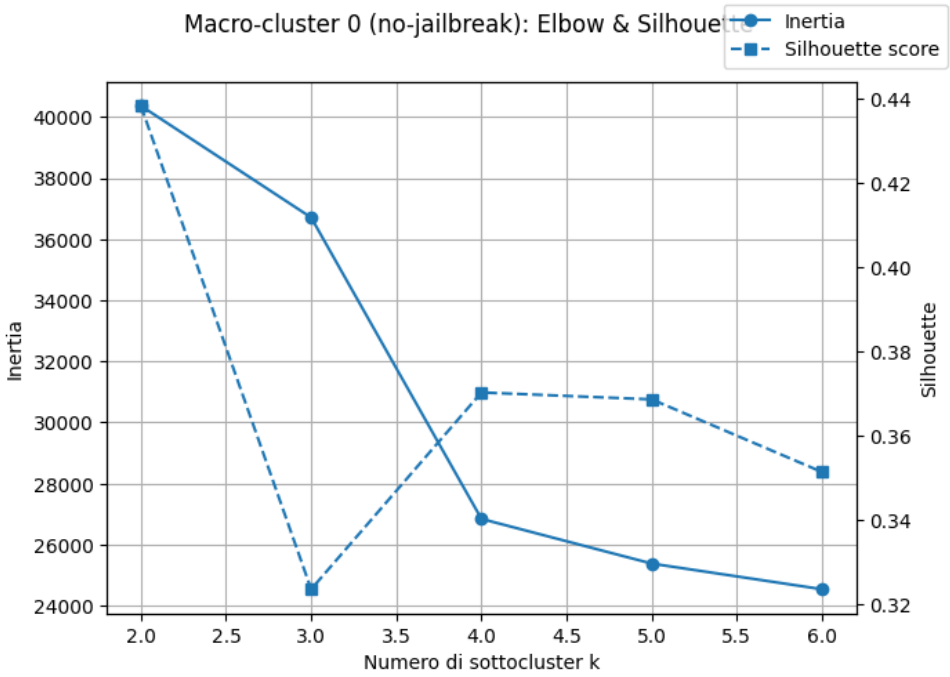

Figure 7: Analisi elbow e silhouette per il Macro-cluster 0 (no-jailbreak) nella Run 2.

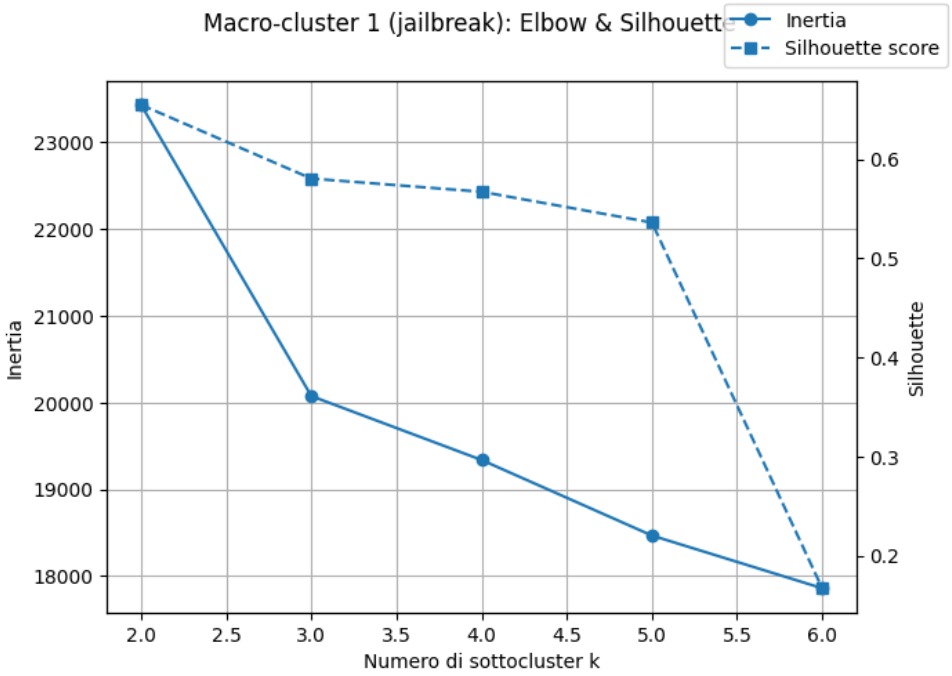

Figure 8: Analisi elbow e silhouette per il Macro-cluster 1 (jailbreak) nella Run 2.

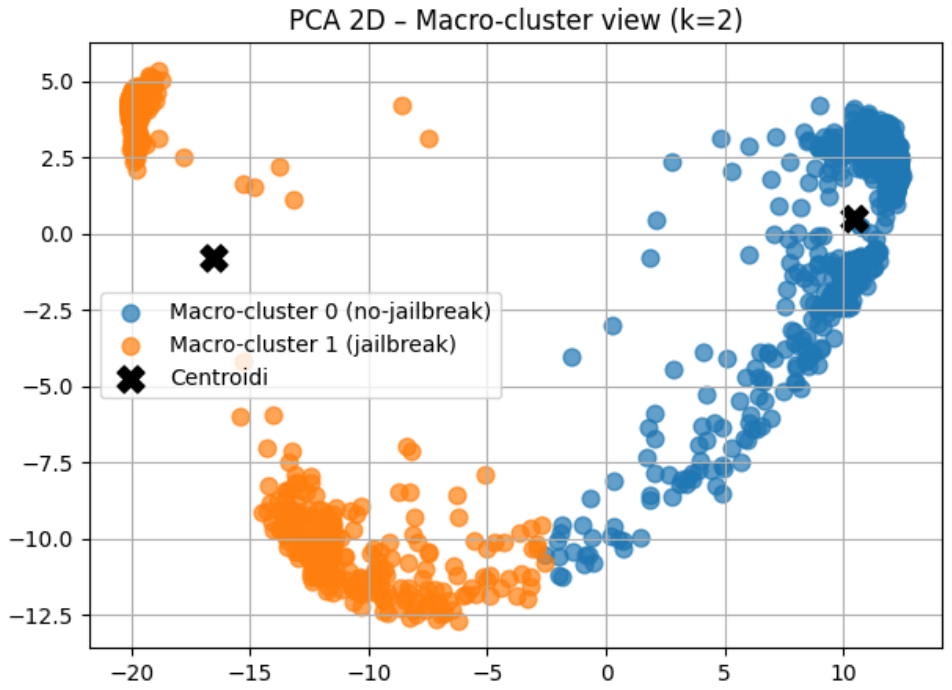

Figure 9: Visualizzazione PCA dei macro-cluster nella Run 2.

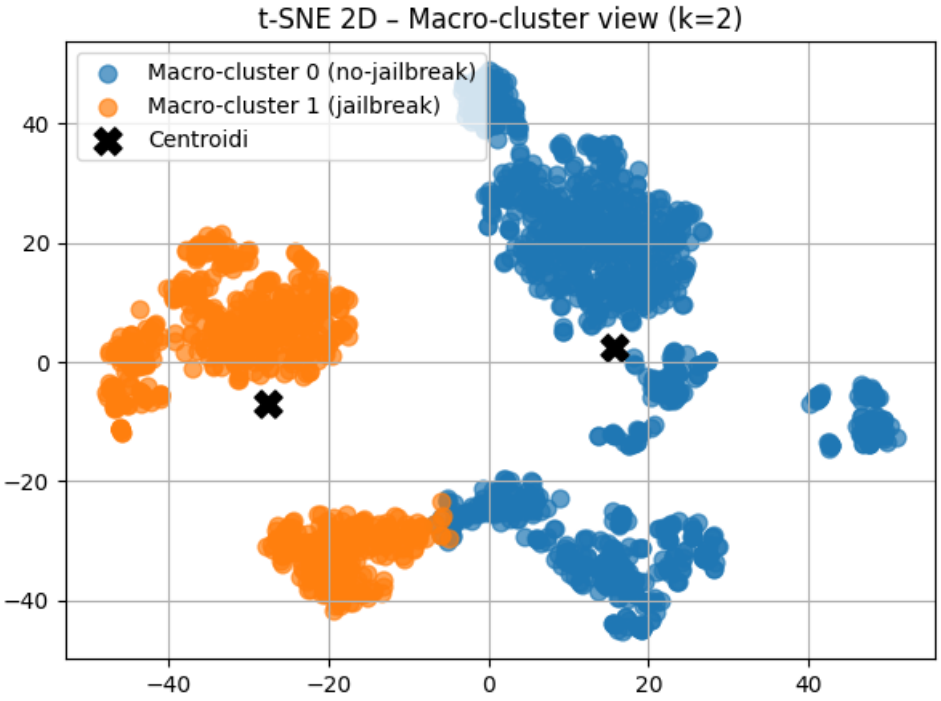

Figure 10: Visualizzazione t-SNE dei macro-cluster nella Run 2.

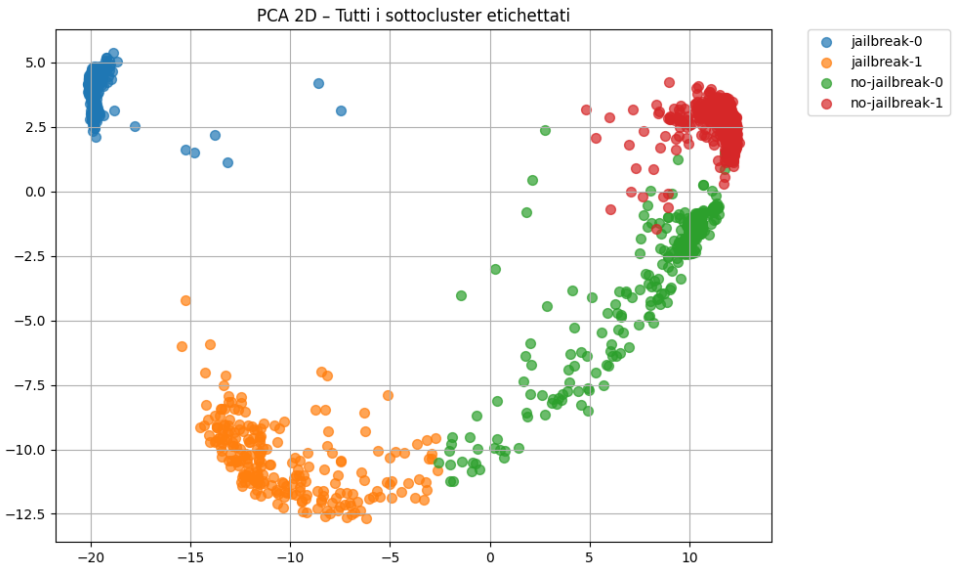

Figure 11: Visualizzazione PCA dei sottocluster all'interno dei macro-cluster nella Run 2.

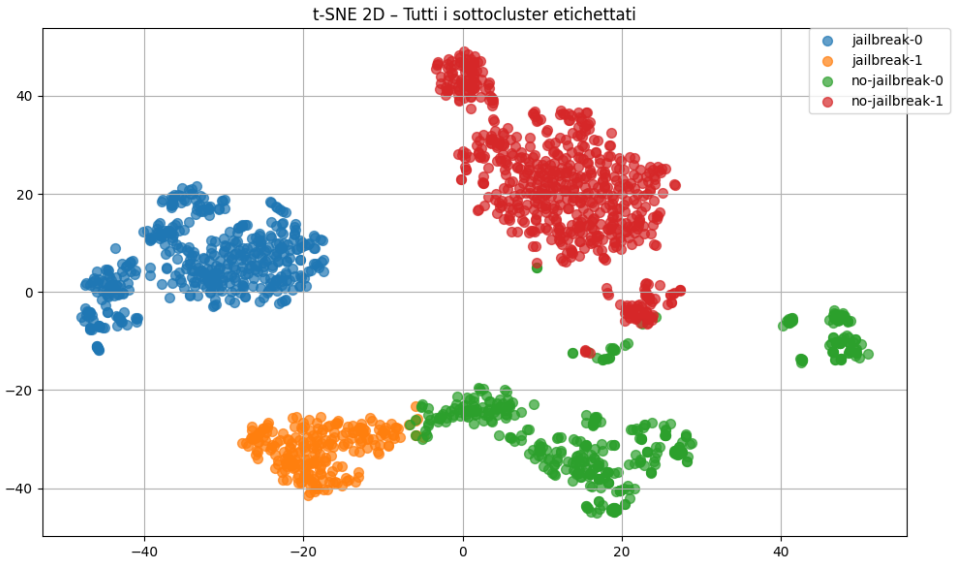

Figure 12: Visualizzazione t-SNE dei sottocluster all'interno dei macro-cluster nella Run 2.