Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

TextEvalScore

Autori

Dott. Luca Esposito

Dott.ssa Onelia Sara Cataldo

Anno Accademico 2024/2025

Indice

Abstract

La valutazione della qualità dei testi generati da modelli di intelligenza artificiale rappresenta una sfida complessa. Recentemente, il framework GPTScore ha introdotto un metodo innovativo per l'analisi multi-aspetto e personalizzata di testi generati, sfruttando le capacità emergenti dei modelli linguistici pre-addestrati. In questo lavoro, presentiamo TextEvalScore, un'applicazione basata su GPTScore che offre un'interfaccia grafica intuitiva per la valutazione automatizzata di dataset testuali. TextEvalScore permette agli utenti di selezionare i prompt di valutazione e i dataset da analizzare, restituendo punteggi dettagliati per diverse metriche qualitative. I risultati sperimentali dimostrano l'efficacia del nostro strumento nel fornire valutazioni coerenti con i giudizi umani e nella personalizzazione delle metriche di scoring. Il nostro contributo mira a rendere più accessibile l'adozione di strumenti di valutazione avanzati per la generazione di testi, riducendo la dipendenza da metriche tradizionali basate su similarità lessicale.

Capitolo 1

Introduzione

Con l'avanzare delle tecnologie di generazione del linguaggio naturale, i modelli pre-addestrati come GPT-3 e Llama-3 hanno raggiunto livelli di qualità testuale straordinari. Tuttavia, la valutazione della qualità di questi output rimane un problema aperto. Metriche convenzionali come BLEU e ROUGE, basate sulla similarità con riferimenti predefiniti, presentano limitazioni nel catturare aspetti più complessi come la coerenza, la factualità e la leggibilità.
Il framework GPTScore ha proposto un approccio innovativo che utilizza modelli linguistici per assegnare punteggi qualitativi ai testi generati, basandosi su istruzioni testuali e apprendimento contestuale. Questo permette un'analisi più flessibile e adattabile rispetto ai metodi tradizionali.
In questo contesto, abbiamo sviluppato **TextEvalScore**, un'applicazione interattiva che implementa il framework GPTScore in un ambiente user-friendly. TextEvalScore consente agli utenti di selezionare il prompt di valutazione, scegliere il dataset di riferimento e ottenere risultati dettagliati sulla qualità del testo generato. Grazie alla sua interfaccia intuitiva e alla capacità di sfruttare modelli pre-addestrati senza necessità di fine-tuning, il nostro strumento rappresenta un passo avanti nella valutazione automatica della qualità del testo.
Nei prossimi capitoli, descriveremo nel dettaglio l'architettura di TextEvalScore, le metodologie adottate per la valutazione e i risultati sperimentali ottenuti.

Capitolo 2

Stato dell'arte

La valutazione automatica della qualità del testo è stata tradizionalmente affrontata con metriche di similarità lessicale, come **BLEU** (Papineni et al., 2002) e **ROUGE** (Lin, 2004), che confrontano il testo generato con un riferimento predefinito. Tuttavia, questi approcci sono limitati nella capacità di catturare aspetti semantici più profondi, come la coerenza e la factualità.
Per superare questi limiti, sono stati introdotti metodi basati su modelli neurali, come **BERTScore** (Zhang et al., 2020) e **MoverScore** (Zhao et al., 2019), che utilizzano rappresentazioni contestuali per valutare la similarità tra il testo generato e il riferimento. Inoltre, metodi come BARTScore (Yuan et al., 2021) sfruttano modelli generativi fine-tunati per assegnare punteggi di qualità ai testi prodotti.
Recentemente, si è assistito all'emergere di approcci basati su modelli pre-addestrati di grandi dimensioni, come **GPT-3** e **LLama-3**, che possono eseguire valutazioni di testo senza necessità di ulteriore addestramento. **GPTScore** (Fu et al., 2024) rappresenta un importante avanzamento in questo campo, permettendo valutazioni multi-aspetto personalizzabili attraverso semplici istruzioni testuali.

Capitolo 3

TextEvalScore

TextEvalScore è un'applicazione progettata per facilitare la valutazione della qualità dei testi generati, combinando la potenza del framework GPTScore con un'interfaccia intuitiva e interattiva.
Resi di facile lettura anche i risultati ottenuti dalla valutazione delle performance e quindi dalla correlazione dei punteggi ottenuti dal modello e da quelli dati dal giudizio umano.

3.0.1 Architettura del sistema

L'applicazione è stata sviluppata in Python (versione 3.11) sfruttando le seguenti librerie:

- tkinter
- openai
- scipy.stats
- sklearn.metrics
- numpy

La struttura del progetto è quella illustrata in Figura 3.1

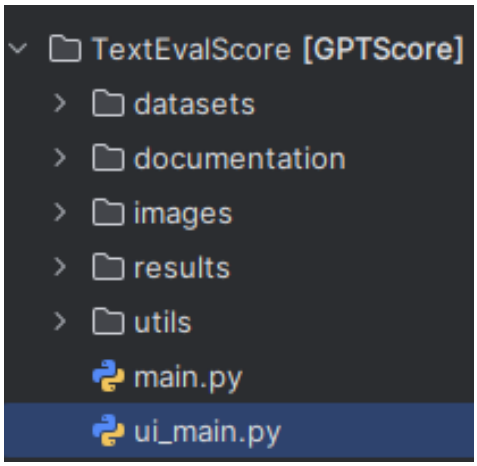

Figura 3.1: Struttura del progetto.

3.0.2 Interfaccia grafica

TextEvalScore dispone di un'interfaccia utente intuitiva che permette di navigare facilmente tra le opzioni di valutazione e visualizzare i risultati in modo chiaro e comprensibile.
All'avvio l'applicazione appare come in figura 3.2

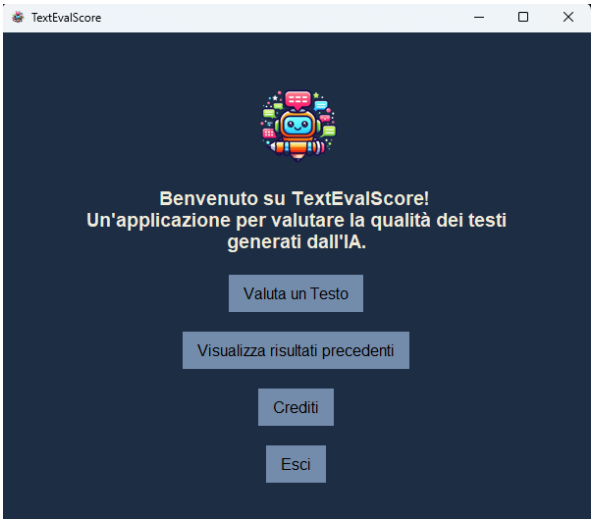

Figura 3.2: Home del progetto.

L'utente ha 4 possibili opzioni:

- Valuta un testo (figura 3.3)
- Visualizza risultati precedenti (figura 3.4 e figura 3.5)
- Crediti (figura 3.6)
- Esci

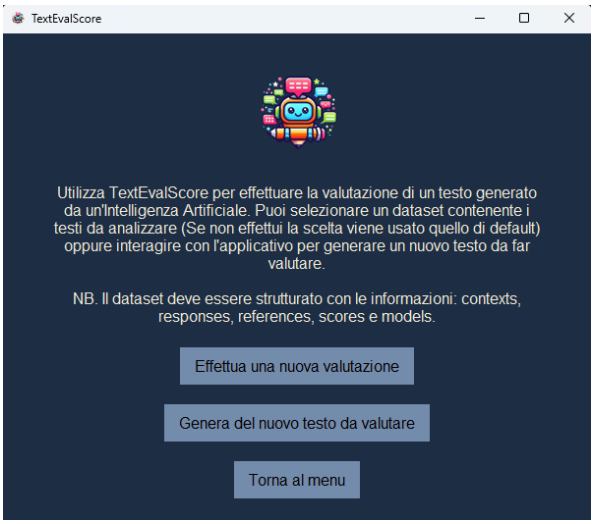

Figura 3.3: Valuta un testo.

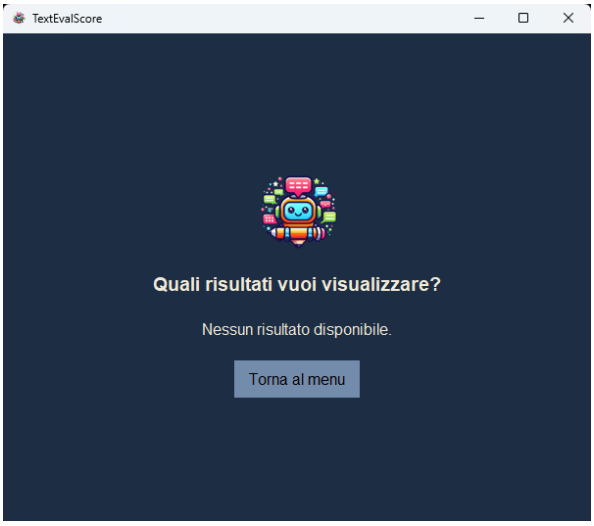

Figura 3.4: Schermata risultati: Nessun risultato disponibile.

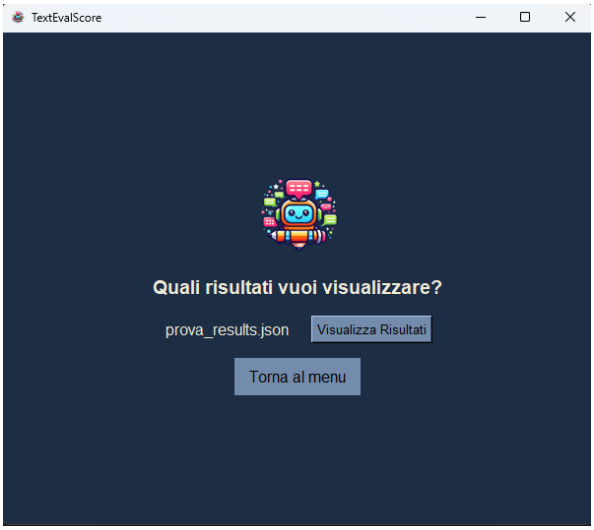

Figura 3.5: Schermata risultati: Lista dei risultati disponibili.

3.0.3 Workflow

Nella schermata iniziale l'utente può selezionare le quattro opzioni mostrate nel paragrafo precedente. Quindi se avviare una nuova valutazione, se visualizzare dei risultati già salvati, oppure visualizzare i crediti dell'applicativo o uscire.
Nel caso in cui si selezioni **Valuta un testo**, all'utente verrà mostrata un'altra finestra (figura 3.3). Qui l'utente potrà decidere se effettuare una nuova valutazione o generare del nuovo testo da poter valutare.
Nel caso in cui venga selezionato **Effettua una nuova valutazione** all'utente verranno mostrate una serie di finestre di dialogo che gli permetteranno di fare delle scelte per la valutazione. La prima (figura 3.7) gli permette di decidere il prompt da utilizzare. Mentre la seconda (figura 3.8) il dataset da utilizzare, quello di default è *dstc9_ data.json*.

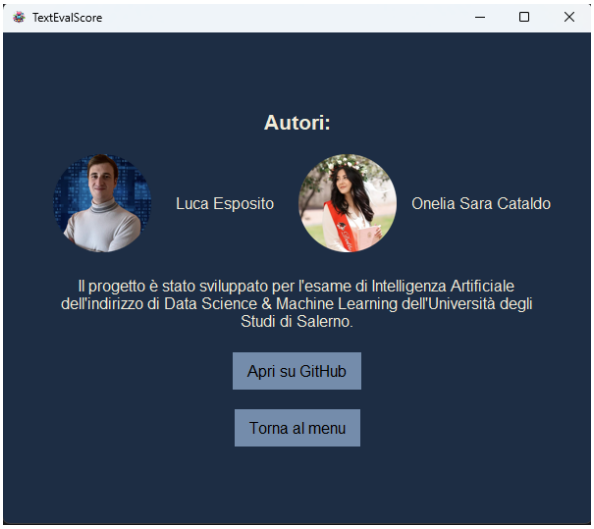

Figura 3.6: Crediti.

Dopo aver effettuato queste scelte partirà l'elaborazione e l'utente sarà aggiornato in real time grazie alla progress bar mostrata in figura 3.9.
Al termine dell'elaborazione verrà mostrata la schermata descritta in figura 3.10.

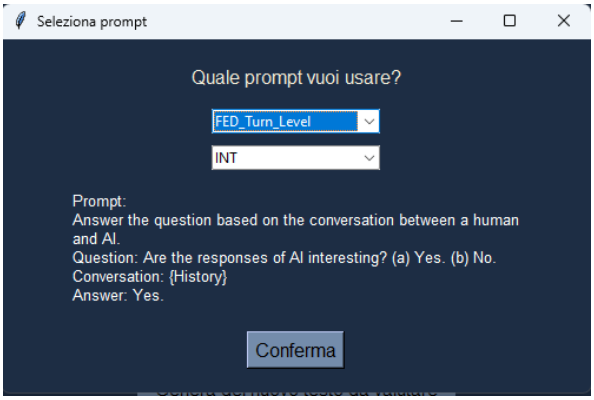

Figura 3.7: Seleziona il promp più adatto.

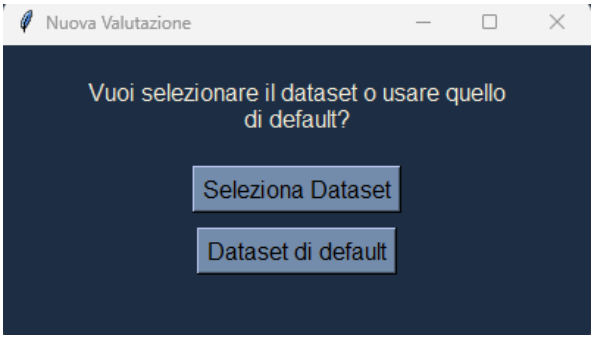

Figura 3.8: Seleziona un dataset da valutare.

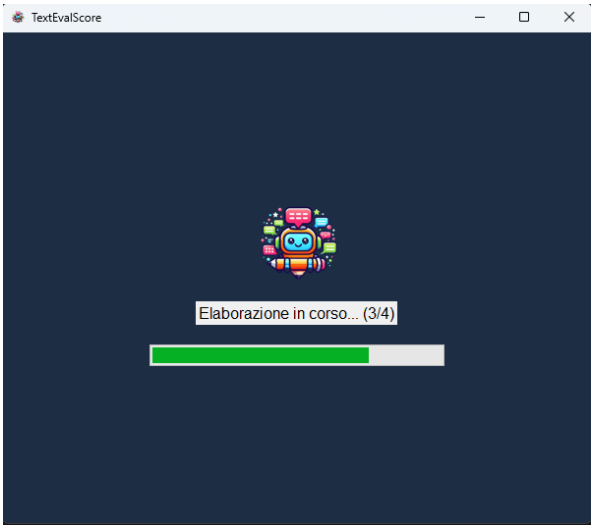

Figura 3.9: Elaborazione in corso.

Nel caso in cui invece venga selezionato **Genera del nuovo testo da valuatare** all'utente verrà mostrata la schermata in figura 3.11. In questa finestra è possibile chattare con il modello in modo tale da generare un nuovo dialogo che verrà salvato in un file .json con la giusta struttura che TextEvalScore si aspetta in modo tale da poter essere utilizzato per una prossima valutazione.
Tornando alla schermata iniziale e cliccando su **Visualizza risultati precedenti** si aprirà la finestra descritta in figura 3.4/3.5. Nel caso in cui siano già state effettuate delle valutazioni sarà disponibile una lista di risultati ottenuti da esse. E cliccando su **Visualizza risultati** dell'elaborazione alla quale siamo interessati possiamo ottenere a video una nuova schermata che ci illustra i risultati ottenuti (figura 3.12).

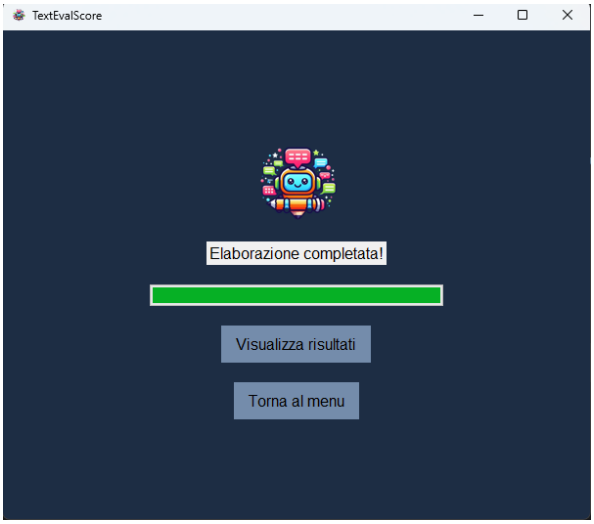

Figura 3.10: Elaborazione completata.

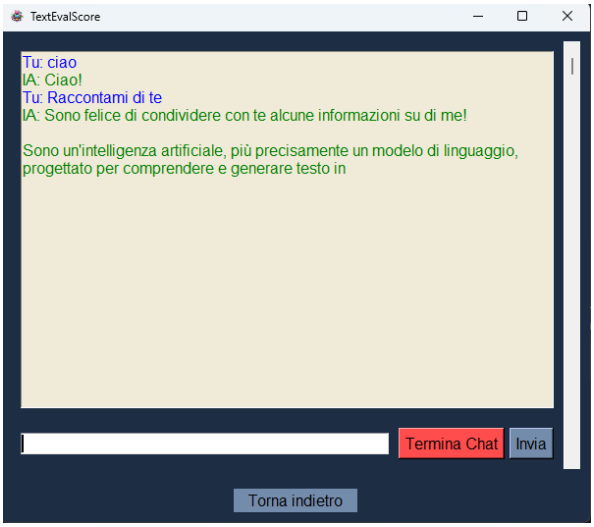

Figura 3.11: Genera un nuovo dialogo.

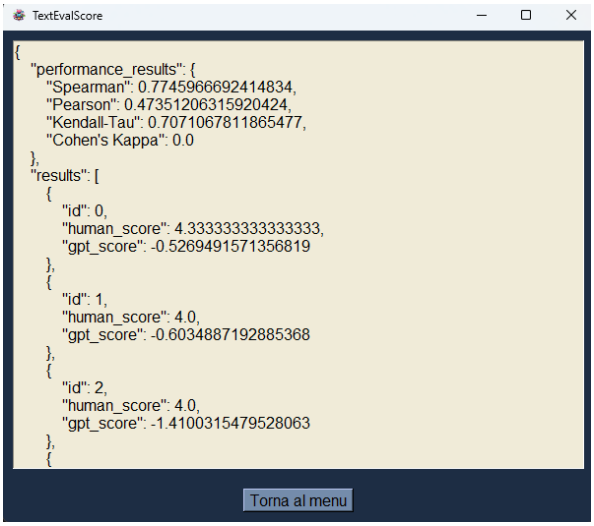

Figura 3.12: Risultati di una specifica elaborazione.

Capitolo 4

Conclusioni

Nel capitolo precedente abbiamo mostrato come vengono visualizzati i risultati ottenuti da una valutazione effettuta con TextEvalScore. Sicuramente l'interfaccia grafica intuitiva ci ha permesso di avere una chiara visione di essi in modo da analizzarli e riutilizzarli per studi futuri e miglioramenti dello stesso applicativo.

4.0.1 Risultati

GPTScore, grazie alla sua flessibilità, ci ha permesso di utilizzare la sua personalizzazione dei prompt per migliorare e personalizzare l'esperienza utente di chi utilizza TextEvalScore. Abbiamo permesso all'utente di selezionare il livello secondo il quale dovrà essere effettuata la valutazione:

- **Turn-Level**: si focalizza sulla risposta più recente
- **Dialogue-Level**: si focalizza sulla coerenza di un'intera conversazione

In base a questa scelta l'utente può anche decidere di utilizzare diversi prompt già impostati in base a dei criteri descritti dal prompt stesso.
Ci siamo preoccupati di testare TextEvalScore con diversi dataset, quello di default e alcuni generati da noi con una delle funzionalità dell'applicativo stesso. Uno dei principali risultati ottenuti e di cui ci siamo accorti è che quando il modello elabora una conversazione in cui i messaggi sono abbastanza brevi e viene effettuata utilizzando un livello di tipo dialogue-level risulta essere più simile al giudizio umano rispetto ad un dialogo più elaborato.
I risultati ottenuti da una valutazione sono stati messi in correlazione con i giudizi umani secondo metriche statistiche, un esempio possiamo vederlo in figura 4.1.

4.0.2 Sviluppi futuri

Sebbene l'esperienza utente è stata curata per garantire quanta più flessibilità a chi utilizza TextEvalScore, ha ancora dei limiti e ci sono caratteristiche che possono essere migliroate in sviluppi futuri allargando le diverse scelte che l'uente può effettuare.
Ne abbiamo individuate alcune:

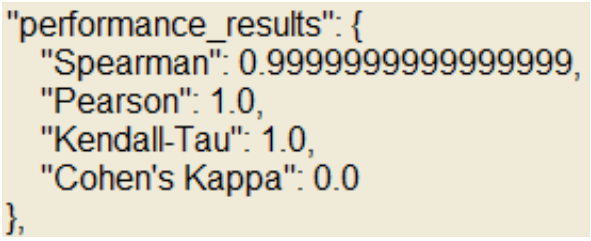

Figura 4.1: Risultati di una correlazione tra scores di gpt e giudizi umani.

- **Scelta del modello da utilizzare**: attualmente l'utente non può decidere con che modello interfacciarsi in fase di valutazione di un dataset perciò uno sviluppo futuro potrebbe includere questa scelta ed eventuali confronti dai risultati ottenuti da modelli diversi.
- **Personalizzare prompt**: attualmente la lista dei prompt che si possono utilizzare è statica e decisa dagli sviluppatori che hanno ripreso quelli utilizzati da GPTScore. Uno sviluppo futuro potrebbe includere la possibilita di far creare un nuovo prompt all'utente rispettando comunque la struttura descritta dal livello selezionato (turn-level/dialogue-level).
- **Valutazione del testo generato**: un altro sviluppo futuro di TextEvalScore è quello di andare a migliorare la valutazione dell'utente del testo generato dall'applicativo rendendo automatico il salvataggio dello stesso nel dataset di output, poichè attualmente avviene manualmente.