

AI e Feedback. Interazione tra agenti umani e artificiali per valutare prove scritte in ambito universitario

AI and Feedback. Assessing written tasks in Higher Education through human and artificial agent interaction

Laura Carlotta Foschi*, Beatrice Doria*, Chiara Laici**, Francesca Gratani**, Laura Screpanti**, Michele Giuliano Fiorentino***, Pier Giuseppe Rossi**, Lorella Giannandrea*, Valentina Grion****, Antonella Montone****

Riassunto

Il feedback automatizzato, reso possibile da strumenti tecnologici avanzati come l'intelligenza artificiale, rappresenta una frontiera emergente per superare alcune delle sfide tradizionali legate alla personalizzazione e alla scalabilità del processo valutativo, soprattutto in classi numerose. In Italia, nonostante l'interesse crescente a livello europeo, l'adozione di valutazioni supportate da tecnologie digitali rimane limitata e presenta numerose sfide. Tali criticità sottolineano l'urgenza di promuovere lo sviluppo professionale dei docenti, attraverso percorsi di formazione mirati a integrare il feedback automatizzato nelle pratiche didattiche, al fine di valorizzare il potenziale trasformativo di questi strumenti. In tale contesto è sorto il PRIN "AI&F", volto a definire una metodologia per l'utilizzo di un framework di *machine learning open-source* per supportare i docenti nel fornire feedback di alta qualità a gruppi numerosi di studenti generando percorsi interattivi e trasformativi in una logica ecosistemica. Il contributo presenta gli attuali avanzamenti della ricerca e delinea le future prospettive di sviluppo.

Parole chiave: feedback; intelligenza artificiale; università; BERT model; LLM

* Università degli Studi di Padova. E-mail: lauracarlotta.foschi@unipd.it.

** Università degli Studi di Macerata. E-mail: chiara.laici@unimc.it.

*** Università degli Studi di Bari Aldo Moro. E-mail: michele.fiorentino@uniba.it.

**** Università Telematica Pegaso. E-mail: valentina.grion@unipegaso.it.

° This paper is the result of a collaborative effort among the authors. In detail: Pier Giuseppe Rossi, Lorella Giannandrea, Valentina Grion and Antonella Montone developed the research framework and supervised the drafting of the paper. Laura Carlotta Foschi is the author of section 2; Beatrice Doria is the author of section 1; Chiara Laici is the author of section 3, Francesca Gratani is the author of the section 3.2; Laura Screpanti is the author of section 3.1; Michele Giuliano Fiorentino is the author of section 4.

Abstract

Automated feedback, made possible by advanced technological tools such as artificial intelligence, represents an emerging frontier to overcome some of the traditional challenges related to the customisation and scalability of the assessment process, especially in large classes. In Italy, despite growing interest at European level, the adoption of digitally supported assessment remains limited and presents numerous challenges. These critical issues underline the urgency of promoting the professional development of teachers through training courses aimed at integrating automated feedback into teaching practice, in order to enhance the transformative potential of these tools. It is in this context that PRIN “AI&F” has emerged, which aims to define a methodology for using an open-source machine learning framework to support teachers in providing high quality feedback to large groups of students, generating interactive and transformative pathways in an ecosystem logic. The paper presents current research progress and outlines future development perspectives.

Key words: feedback; artificial intelligence; higher education; BERT model; LLM

First submission: 04/12/2024, accepted: 13/12/2024

1. Introduzione

Il feedback costituisce un elemento cruciale per il miglioramento dell'apprendimento, poiché consente agli studenti di confrontare le proprie prestazioni con standard di riferimento al fine di ridurre il divario tra risultati attuali e obiettivi attesi (Nicol, 2010; Price et al., 2010). La ricerca educativa, inizialmente ancorata a modelli trasmissivi, ha progressivamente adottato un approccio socio-costruttivista, concependo il feedback come un processo dialogico che promuove l'attiva costruzione di conoscenze da parte degli studenti (Nicol, 2018). Questo cambiamento paradigmatico ha posto maggiore enfasi sul ruolo dello studente come agente attivo, in grado di decodificare, analizzare e integrare le informazioni ricevute per perfezionare il proprio apprendimento (Grion e Serbati, 2019; Grion et al., 2021; Nicol, 2021).

In questo contesto, il feedback automatizzato, reso possibile da strumenti tecnologici avanzati come l'intelligenza artificiale (AI), rappresenta una frontiera emergente per superare alcune delle sfide tradizionali legate alla personalizzazione e alla scalabilità del processo valutativo, soprattutto in classi numerose (Grion e Cesareni, 2016). Tali sistemi, progettati per fornire un feedback

immediato e dettagliato, offrono agli studenti l'opportunità di riflettere autonomamente sulle proprie prestazioni e identificare le aree di miglioramento (Steiss et al., 2024). Inoltre, essi possono agevolare l'adozione di pratiche valutative innovative, come quelle basate su valutazioni formative e partecipative, in grado di coinvolgere attivamente gli studenti nel processo di apprendimento (Redecker e Punie, 2017).

Nonostante le potenzialità del feedback automatizzato, la letteratura sottolinea la necessità di una sua integrazione pedagogicamente consapevole (Raffaghelli, 2024). Affinché tali sistemi possano supportare realmente l'apprendimento, è cruciale che docenti e studenti acquisiscano competenze adeguate per interagire con questi strumenti. In particolare, i docenti devono ripensare i propri approcci didattici, adottando strategie che coniughino efficacemente tecnologie digitali e processi di apprendimento attivo, mentre gli studenti necessitano di sviluppare capacità di analisi critica per decodificare e applicare il feedback ricevuto (Doria et al., 2024; Giannandrea, 2019; Raffaghelli, 2024).

Tuttavia, la diffusione di tali strumenti non è priva di criticità. Gli studi evidenziano che, se non correttamente progettati, i sistemi automatizzati possono produrre feedback poco chiari o persino controproducenti, richiedendo un'attenta valutazione dell'efficacia delle interazioni uomo-macchina (Facer e Selwyn, 2021). Inoltre, le tecnologie basate sull'AI sollevano questioni etiche e pedagogiche, come l'equità nell'accesso, la privacy dei dati e la sostenibilità dell'integrazione digitale nei contesti educativi (Raffaghelli, 2024).

In Italia, nonostante l'interesse crescente a livello europeo per l'adozione di valutazioni supportate da tecnologie digitali, il loro utilizzo rimane limitato (Doria e Picasso, 2024). Tale divario sottolinea l'urgenza di promuovere lo sviluppo professionale dei docenti, attraverso percorsi di formazione mirati a integrare il feedback automatizzato nelle pratiche didattiche, al fine di valorizzare il potenziale trasformativo di questi strumenti. In questo contesto il Progetto di Rilevante Interesse Nazionale (PRIN) "AI&F", coinvolgente le Università di Padova (UniPD), Macerata (UniMC) e Bari Aldo Moro (UniBA), si propone di definire una metodologia per l'utilizzo di un framework di *Machine Learning* (ML) open-source per supportare i docenti nel fornire feedback di alta qualità a gruppi numerosi di studenti, generando percorsi interattivi e trasformativi in una logica ecosistemica. Il progetto si struttura nei seguenti Work Packages (WP): WP1 (UniPD) Analisi, WP2 (UniMC) Sviluppo, WP3 (UniBA) Piloting.

2. WP1: Obiettivi, metodologia, risultati

Il WP1 ha previsto due nuclei di ricerca (*Fig. 1*) finalizzati ad analizzare nelle tre Università coinvolte nel PRIN: 1) le pratiche di feedback e di feedback

automatizzato messe in atto dai docenti nei loro insegnamenti e 2) le percezioni ed esperienze d'uso degli studenti universitari rispetto all'utilizzo del feedback e del feedback automatizzato. In relazione al primo nucleo, sono state dapprima analizzate le pratiche di feedback dichiarate nei Syllabi e successivamente si è approfondito come il feedback, incluso quello automatizzato, viene definito e utilizzato dai docenti attraverso la conduzione di tre focus group. In relazione al secondo nucleo si è proceduto con la costruzione e somministrazione di un questionario agli studenti volto a indagare le loro percezioni circa l'utilità di specifiche situazioni e strumenti di feedback e di feedback automatizzato e le loro esperienze in merito.

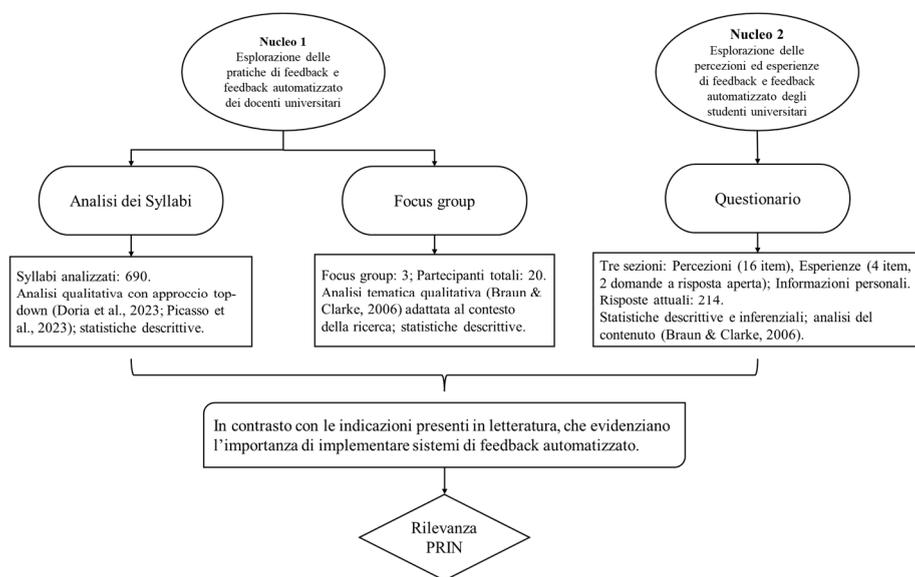


Fig. 1 - Flusso di lavoro del WP1

Per indagare le pratiche di feedback (1a), sono stati analizzati i Syllabi di un campione rappresentativo di docenti appartenenti alle tre università partner (UniPD: 322; UniMC: 150; UniBA: 290), selezionato stratificando l'intera popolazione docente per i quattordici settori scientifico-disciplinari. I Syllabi completi analizzabili sono risultati 690. Su questi è stata condotta un'analisi qualitativa con approccio top-down (Doria et al., 2023; Picasso et al., 2023), i cui risultati sono stati poi elaborati mediante analisi statistiche descrittive.

Per approfondire le pratiche di feedback (1b), sono stati condotti tre focus group in tre diversi dipartimenti, uno per ciascun Ateneo, con un campione di 20 docenti (6-8 per ateneo). I focus group hanno esplorato i macrotemi del feed-

back (definizione, modalità di utilizzo, tipologie utilizzate) e del feedback automatizzato (utilizzo, vantaggi, criticità). Le trascrizioni sono state analizzate qualitativamente adattando al contesto della ricerca i passaggi indicati da Braun e Clarke (2006), con successiva quantificazione delle occorrenze per ciascuna categoria.

Per indagare le percezioni e le esperienze degli studenti relativamente alle pratiche di feedback (2), è stato progettato un questionario composto da tre sezioni. La prima, articolata in sedici item con scala di risposta di accordo da 1 a 7, mirava a indagare le percezioni in merito a specifiche situazioni e strumenti di feedback e di feedback automatizzato. Le situazioni si rifacevano alle tre fonti primarie di feedback indicate da Panadero e Lipnevich (2022): docente, compagni, computer. Gli strumenti sono stati scelti tra quelli più utilizzati nel contesto italiano (e.g., ChatGPT, Grammarly). La seconda sezione si proponeva di esplorare, attraverso quattro item (opzioni di risposta: Mai, Qualche volta, Spesso, Sempre), la frequenza con cui gli studenti hanno sperimentato pratiche di feedback e di feedback automatizzato nei loro corsi universitari. Comprende inoltre due domande aperte volte a esplorare la migliore e la peggiore esperienza di feedback. La terza era infine deputata alla raccolta di informazioni personali. Attualmente sono state registrate 214 risposte. I dati degli item sono stati elaborati mediante statistiche descrittive e inferenziali e i dati qualitativi sottoposti ad analisi del contenuto (Braun e Clarke, 2006).

Vengono di seguito brevemente riportati i risultati relativi a quanto emerso in relazione al feedback automatizzato. L'analisi dei Syllabi (1a) ha evidenziato che solo il 9.71% (n = 67) dei docenti dichiara di utilizzare nei propri insegnamenti pratiche di feedback, con un'adozione praticamente inesistente (n = 1; .14%) del feedback automatizzato (nell'accezione di *Computer Based Assessment practices*, Tonelli et al., 2018). Questo risultato è in linea con l'analisi dei Syllabi condotta a livello italiano da Doria e Picasso (2024), che ha rilevato un utilizzo molto limitato di pratiche di valutazione e feedback supportate dalla tecnologia.

La limitata adozione del feedback automatizzato è emersa anche dall'analisi dei focus group (1b). Su un totale di 205 occorrenze relative alle due macrocategorie, *Feedback* (n = 159, 77.58%) e *Feedback automatizzato* (n = 46, 22.43%), la categoria "Uso del feedback automatizzato" rappresenta solo il 10.24% delle quotazioni (n = 21). Di queste ultime, inoltre, unicamente 15 (7.31%) si riferiscono all'utilizzo del feedback automatizzato, mentre le restanti 6 (2.93%) ne evidenziano il non uso o la non conoscenza¹.

¹ Per i risultati completi relativi all'analisi dei Syllabi e dei focus group si rimanda a Doria et al. (*in press*).

Infine, dal questionario somministrato agli studenti (2) è emerso che tutte le otto situazioni e strumenti di feedback automatizzato proposti sono percepiti come molto utili: la media delle risposte di tutti gli item è superiore al punto medio della scala (4). Le medie variano da 4.71 (DS: 1.40) per l'item S11 (Feedback ottenuto da conversazioni con chatbot come ChatGPT) a 5.46 (DS: 1.19) e 5.45 (DS: 1.25) rispettivamente per gli item S9 (Feedback automatizzati virtuali) e S14 (Feedback da sistemi di analisi del testo). Tuttavia, nonostante l'elevata utilità percepita, l'esperienza effettiva degli studenti con questa tipologia di feedback in contesti universitari è limitata. In particolare, riguardo all'utilizzo di sistemi digitali con forme di feedback automatizzati, la maggior parte degli studenti ha dichiarato di aver frequentato corsi in cui sono stati impiegati tali sistemi solo "Qualche volta" (55.11%) o addirittura "Mai" (23.8%), mentre solo una minoranza "Spesso" (20.1%) e quasi nessuno "Sempre" (.9%).

I risultati presentati sono in linea con la letteratura, che evidenzia una prevalenza di approcci valutativi tradizionali e una certa resistenza all'adozione di nuove tecnologie da parte dei docenti universitari (e.g., Doria e Picasso, 2024; Panadero et al., 2019). Tuttavia, questi risultati contrastano con le indicazioni che sottolineano l'importanza di implementare sistemi di feedback automatizzato per rilevare il reale processo e progresso di apprendimento degli studenti e, soprattutto, fornir loro feedback tempestivi e personalizzati affinché possano migliorare le proprie strategie di apprendimento (e.g., Boud et al., 2018). In tale contesto, il PRIN assume particolare rilevanza, proponendosi di sviluppare un metodo, un'interfaccia *user-friendly* e tutorial per l'utilizzo di framework di ML, finalizzati a generare feedback di qualità per supportare un apprendimento efficace, specialmente in gruppi numerosi di studenti, rispondendo così alla necessità evidenziata.

3. WP2: Sviluppo della sperimentazione

La ricerca sul feedback nell'*higher education* ha visto negli ultimi anni l'affermarsi di una svolta paradigmatica (Winstone e Carless, 2019), passando da un feedback prevalentemente inteso come trasmissione unidirezionale di un commento dal docente agli studenti (Sadler, 2010) a un feedback *learning centered* (Laici e Pentucci, 2023). Quest'ultimo enfatizza il processo di apprendimento degli studenti attraverso un coinvolgimento attivo e, a prescindere dalla fonte del feedback, pone attenzione agli aspetti di interazione tra i vari attori coinvolti quali docenti, studenti, pari ed anche strumenti tecnologici (Nicol e Macfarlane-Dick, 2006; Lipnevich e Panadero, 2021; Laici, 2021; Giannandrea et al., 2024). Oggi possiamo, infatti, confrontarci con diverse tipologie di processi di feedback che possono coinvolgere agenti umani (AU) ma anche agenti

artificiali (AA) intelligenti che abilitano la restituzione di un feedback automatizzato (AA), in una logica non esclusiva ma integrata.

La prospettiva è quella che posiziona l'AI come un partner, come facilitatore e co-autore nei contesti educativi (Panciroli e Rivoltella, 2023), poiché non vi è una totale delega dei processi di feedback ai sistemi automatizzati, ma una collaborazione alla progettazione di percorsi interattivi e trasformativi in una logica ecosistemica (Gupta et al., 2024) dove gli attori interagiscono, si attivano e agiscono sui commenti di feedback (Nieminen e Carless, 2023).

Il feedback automatizzato può contribuire quindi in diverse direzioni in coerenza con un approccio di feedback attivo e generativo (Rossi et al. 2018) supportando i docenti nel fornire a un ampio numero di studenti un feedback tempestivo, utilizzabile quando effettivamente a loro serve, per migliorare il proprio apprendimento e quindi in itinere e in modo non occasionale. Questo consentirebbe di rendere il processo realmente interattivo implementando anche diversi livelli di coinvolgimento come, ad esempio, un feedback iterativo e processi di *feedback loop*. Inoltre, potrebbe essere un supporto per valorizzare il processo di *inner feedback* (Tam, 2024) e per affinare le competenze valutative e il giudizio valutativo degli studenti (Bearman et al., 2024).

Per intraprendere questo percorso di ricerca ci si è chiesti quindi come i sistemi di AI potessero supportare il docente nell'analizzare un artefatto degli studenti e fornire quindi un feedback tempestivo.

La sfida di riuscire a fornire un feedback tempestivo è ancora più complessa quando le classi sono molto numerose e variegate (Raffaghelli et al., 2018; Rannieri et al., 2019). Una soluzione spesso adottata è l'utilizzo di prove che permettano la correzione automatica (ad esempio i quiz a risposta chiusa), ma una tale scelta limita le tipologie di prove e di competenze indagate.

Nel progetto PRIN "AI&F" sono state sperimentate due strategie adatte per due situazioni diverse. La prima categoria è rivolta alla correzione di risposte libere su specifici contenuti: le strutture linguistiche sono molteplici, ma i significati sono definiti. La seconda è relativa a elaborati la cui struttura è data, ma i contenuti sono scelti da ogni singolo studente. Appartengono a questa categoria progetti, compiti autentici aperti, analisi di contesti.

Per la valutazione delle risposte dell'AI sono state utilizzate valutazioni triangolate effettuate da tre o più docenti che prima hanno corretto separatamente in base a una rubrica condivisa e poi si sono confrontati sulle situazioni non coincidenti. Si precisa che anche in tal caso e pur utilizzando una valutazione a 4 livelli, si avevano situazioni tra il 5% e il 10% di non completa condivisione tra i tre correttori. Si precisa, inoltre, che il confronto tra i risultati della valutazione triangolata e quella effettuata da vari docenti singoli ha mostrato uno scarto tra il 20% e il 40%.

3.1 Tipologia A: correzione di risposte libere su specifici contenuti

Ci riferiamo alla correzione di elaborati della lunghezza media di 400 caratteri, quindi risposte brevi a singole domande, dell'ordine di poche righe.

Per la classificazione delle risposte aperte è stata utilizzata una variante del modello BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2018). Rispetto ai modelli multilingue o general-purpose, il modello selezionato (*dbmdz/bert-base-italian-cased*), addestrato su un ampio corpus di testi italiani, offre una comprensione più accurata delle risposte aperte in italiano, garantendo migliori prestazioni nel contesto educativo, dove le risposte possono contenere frasi non standard o termini tecnici. Ogni risposta è stata tokenizzata utilizzando il tokenizer di BERT, che divide il testo in parole o sottoparole e per ogni token della risposta sono stati calcolati vettori numerici densi e ad alta dimensione, gli *embeddings*. Lo stesso processo è stato applicato ai testi target (TT), permettendo così il confronto tra i TT e le risposte degli studenti.

Questa soluzione usa tecnologie del ML. Analizzare la risposta dello studente con sistemi di ML (senza adattare preventivamente il sistema al compito proposto) fornisce risultati generici, spesso poco rispondenti alle prospettive con cui il docente ha elaborato la domanda. Anche impiegando testi specifici per l'adattamento del modello al compito proposto, i risultati non sono stati soddisfacenti.

Abbiamo allora chiesto all'AA di confrontare le risposte degli studenti con dei TT elaborati in anticipo dai docenti o estratti dalla analisi di alcune risposte degli studenti. Il confronto tra l'intera risposta dello studente con un unico testo target si rivela non del tutto affidabile (successo dell'ordine del 60%-70%). Risultati dell'ordine del 90% sono stati ottenuti individuando sottotemi che dovrebbero essere contenuti nella risposta, seppur breve, elaborando per ogni sottotema una serie di brevi TT non equivoci e confrontando il testo dello studente separatamente con ogni serie. L'AA fornisce per ogni confronto la similarità ovvero un numero tra 0 e 1 che indica la distanza del testo dalla serie dei TT. Il docente, analizzando alcuni casi, decide dove porre l'asticella per ogni serie ovvero il valore di similarità al di sotto del quale si ritiene non accettabile la risposta dello studente. Si assegna poi il valore 1 a tutte le risposte superiori e 0 a quelle inferiori. Sicuramente alcuni testi validi degli studenti saranno sotto tale livello come alcuni testi non validi saranno al di sopra. Per tener conto di tale elemento in alcuni casi si è preferito assegnare 0,5 punti agli elaborati che si collocavano in prossimità dei valori soglia. Successivamente, si assegna il peso che ogni serie ha sulla valutazione complessiva, connesso alla rilevanza che essa ha nella risposta, e il valore iniziale di ogni serie è moltiplicato per il

peso. La valutazione complessiva per il singolo studente si ottiene sommando il valore ottenuto per ogni serie.

Si può, infine, mostrare allo studente la valutazione ottenuta per ogni sotto-tema e questo permette di fornire un feedback puntuale. Abbiamo attivato un sistema di chat asincrona e lo studente può indicare situazioni in cui non concorda con la valutazione dell'AA. Riteniamo, infatti, che le valutazioni fornite dall'AA possano in qualche caso non essere coerenti e rispondenti all'effettivo livello della prova dello studente, per quanto casi simili siano stati riscontrati anche nelle correzioni dell'AU in misura considerevole.

Per far fronte a questi casi di disallineamento, più che perfezionare il sistema informatico, sembra rilevante intervenire ibridando percorsi analogici e digitali, ovvero attivando una triangolazione tra studente, AU (docente) e AA. Il docente può infatti restituire una risposta alla segnalazione dello studente in cui fornisce il suo parere, favorevole o problematizzante, rispetto alle osservazioni dello studente, avviando così un ciclo di feedback utile non solo alla consapevolezza della valutazione ottenuta, ma anche al miglioramento del processo di apprendimento.

Un ruolo chiave in tale procedura è assunto dalla predisposizione dei TT; tanto più i TT saranno ben costruiti, tanto più la classificazione dell'AA sarà affidabile ed efficace. Infatti, una causa di inaffidabilità del testo target è spesso legata alla presenza di contenuti non direttamente connessi al sottotema.

Il processo proposto permette di predisporre in breve tempo un sistema di valutazione fortemente coerente con le richieste del docente. Chiaramente ogni prova richiede di elaborare specifici TT.

3.2 Tipologia B: correzione di elaborati aperti (ad esempio progetti)

Il secondo caso si riferisce a elaborati complessi, anche oltre le 5000 battute, in cui la struttura e la logica potrebbero essere costanti, mentre il contenuto specifico potrebbe essere scelto dal singolo studente arbitrariamente. Sono spesso tipologie di elaborati frequenti in specifici insegnamenti o anche in corsi di laurea. Si pensi, nel caso di Scienze della formazione primaria, alla progettazione di sessioni didattiche. In tal caso sebbene la struttura e la logica progettuale siano dati, ogni elaborato potrebbe affrontare contenuti e argomenti differenti, come differenti potrebbero essere obiettivi e finalità.

In tali casi ci siamo mossi con sistemi di Generative AI (GenAI). Anche in questo caso abbiamo inizialmente effettuato una correzione manuale di elaborati di una certa tipologia (ed. esempio progettazione didattica di una sessione in base a un modello specifico e per una specifica disciplina e per uno specifico anno di corso). Abbiamo poi esplicitato i criteri adottati durante la valutazione, non solo elaborando una rubrica, ma anche analizzando passo passo le logiche

adottate dal valutatore manuale e i passaggi effettuati durante la correzione. Nel caso specifico l'AA invece di analizzare la similarità tra testo dello studente e testo target, analizza la coerenza, valuta la sostenibilità, ecc.

Nella ricerca, un primo passo è stato effettuato sottoponendo l'intero elaborato a un sistema di GenAI, come ChatGPT-4, avendo precisato i criteri di massima adottati. Pur "conoscendo" il sistema e le logiche progettuali, che ha esplicitato abbastanza correttamente, i feedback forniti nella correzione del singolo elaborato dello studente in tale modalità sono risultati generici. Si è passati pertanto anche in questo caso a suddividere il percorso in sotto-aspetti da analizzare e si è lavorato sul *prompt engineering* per costruire una serie di domande che riprendessero il processo utilizzato dall'AU. Anche in questo caso la parcellizzazione del percorso permette di ottenere una maggiore affidabilità probabilistica.

Il passaggio successivo consisterà nell'ingegnerizzare il processo, ovvero costruire una struttura unica che permetta di sottoporre l'elaborato dello studente a tutta la successione dei prompt.

In questa seconda tipologia la costruzione del processo richiede dei tempi significativi, giustificabili solo se la prova esaminata non avviene una tantum, ma è frequente nell'insegnamento o nel corso degli studi.

4. Scenari futuri

L'utilizzo di un sistema automatizzato, coadiuvato dalla AI, per la clusterizzazione delle risposte, la correzione delle stesse, l'invio di feedback e la valutazione di compiti, richiede uno studio delle potenzialità che tale sistema deve possedere, in riferimento agli aspetti specifici delle diverse discipline. Per rispondere adeguatamente alle esigenze educative, pedagogiche e disciplinari esso dovrà svilupparsi in modo da affrontare una serie di sfide che possano garantire un'interazione dinamica e costruttiva tra studente e docente, attraverso l'invio di feedback e una valutazione complessiva accurata.

In tale scenario, sembra essere necessario caratterizzare l'utilizzo del sistema automatizzato, a seconda della disciplina che lo utilizza. In particolare, ci si interroga su quanto la specificità disciplinare (per es. in discipline differenti come matematica e pedagogia) possa influire sull'assetto del sistema stesso.

I compiti autentici, utilizzati ad esempio in ambito matematico, sono caratterizzati da soluzioni complesse da diversi punti di vista: la varietà delle strategie risolutive, il linguaggio altamente specializzato e l'utilizzo di rappresentazioni grafiche, geometriche e algebriche.

La pluralità di approcci possibili risolutivi ad un medesimo problema rende necessaria la loro previsione e la loro analisi, per consentirne una classificazione completa ed esaustiva. I modelli di ML, grazie alla loro capacità di apprendere da grandi volumi di dati, dovranno evolvere dinamicamente per includere approcci nuovi e non convenzionali, migliorando la capacità di riconoscere soluzioni valide anche se non allineate perfettamente ai modelli predefiniti.

Lo studente, infatti, può scegliere e applicare diverse strategie risolutive utilizzando varie rappresentazioni. Pertanto, è necessario educare il sistema automatizzato a riconoscere la logica sottostante e valutare la chiarezza, la correttezza e la completezza del ragionamento effettuato (Fiorentino et al., 2024).

Gli sforzi fino ad ora compiuti con l'obiettivo di migliorare la pratica educativa, attraverso questi nuovi mezzi, sono giustificati dagli innumerevoli vantaggi che essi offrono sia ai docenti sia agli studenti.

I vantaggi per il docente riguardano dapprima la rapidità di restituzione dei feedback agli studenti con conseguente diminuzione delle tempistiche di revisione degli elaborati in classi molto numerose.

In questo modo viene ridotto notevolmente il carico di lavoro del docente che potrà utilizzare il tempo recuperato per progettare attività creative e interattive in aula, incrementando il rapporto diretto con lo studente e garantendo un maggiore supporto individuale. Infatti, un importante strumento che il sistema fornisce è una sezione dedicata all'interazione diretta tra studente e docente. Dopo aver ottenuto la valutazione presentata attraverso la tabulazione dei diversi tipi di errore, il docente, può rispondere facilmente e tempestivamente, offrendo così un feedback mirato e personalizzato. Inoltre, in caso di insuccesso, tale procedura può essere iterata fornendo ulteriori spiegazioni a eventuali richieste.

Una delle maggiori difficoltà affrontate dai docenti è la valorizzazione della valutazione nelle discipline, in ottica formativa: l'utilizzo di un sistema di ML garantisce una valutazione in grado di restituire feedback migliorativi ad un grande numero di studenti, rispondendo ad una elevata varietà di proposte risolutive. Questo riduce il rischio di pregiudizi o discrepanze tra le valutazioni, offrendo una maggiore equità nel processo di costruzione delle soluzioni.

D'altra parte, l'introduzione di queste nuove modalità di valutazione offre vantaggi anche per lo studente. Come già affermato in precedenza, la diminuzione dei tempi di correzione permette la tempestiva ricezione dei feedback, i quali saranno specifici, individualizzati e forniranno suggerimenti concreti per migliorare la capacità di autoregolazione e autovalutazione degli studenti. Affinché una valutazione sia efficace, ossia comporti un miglioramento effettivo del processo di insegnamento e apprendimento, questa deve essere chiara, comprensibile e deve consentire l'identificazione dei punti di forza e degli elementi

migliorativi per ogni singolo studente. La precisione del feedback riduce le ambiguità e le incertezze che spesso accompagnano le valutazioni tradizionali. L'esplicitazione dei criteri valutativi, infatti, garantisce un alto livello di trasparenza.

In merito a quanto detto riguardo la comunicazione diretta tra docente e studente, la possibilità di chiedere chiarimenti quasi in tempo reale attraverso il dispositivo, concede un'opportunità di interazione con il docente che può rispondere con puntualità a domande specifiche, migliorando il processo di apprendimento, laddove risulta fondamentale un feedback immediato per evolvere verso la soluzione corretta.

L'evoluzione della tecnologia offrirà strumenti sempre più sofisticati e sensibili alle diversità di approccio degli studenti che potranno trarre grandi benefici da queste tecnologie. In tal modo si migliora non solo l'efficienza della valutazione, ma anche la qualità dell'interazione tra docenti e studenti.

Inoltre, tra gli scenari futuri, si prevede di fornire all'AI una varietà di compiti che prevedono, oltre al problema aperto, compiti progettuali caratteristici di differenti discipline. In tale prospettiva sarà necessario strutturare un sistema automatizzato multi-agente, in grado di interagire in maniera differente a seconda degli aspetti strutturanti delle varie discipline. Ad esempio, facendo riferimento all'ambito matematico-scientifico, i modelli di ML dovranno essere addestrati per riconoscere la sintassi specifica e identificare errori comuni o scritture ambigue. Inoltre, il sistema dovrà essere in grado di distinguere fra diverse convenzioni notazionali e adattarsi alla specifica modalità di presentazione scelta dallo studente.

Bisognerà inoltre, predisporre un ambiente digitale in cui gli studenti possano risolvere problemi disciplinari specifici attraverso strumenti di scrittura digitale che permettano l'inserimento e la successiva lettura delle diverse soluzioni.

References

- Bearman M., Tai J., Dawson P., Boud D., and Ajjawi R. (2024). Sviluppo del giudizio valutativo per un'epoca di intelligenza artificiale generativa. *Assessment & Evaluation in Higher Education*, 49(6): 893-905. DOI: 10.1080/02602938.2024.2335321.
- Boud D., Ajjawi R., Dawson P., and Tai J. (2018). *Developing evaluative judgement in higher education*. London: Routledge.
- Braun V. e Clarke V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77-101.
- Devlin J., Chang M.-W., Lee K., and Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 1810.04805. DOI: 10.48550/arXiv.1810.04805.

- Doria B., Foschi L.C., Slaviero G., Zaggia C., and Grion V. (*in press*). Feedback e feedback automatizzato: Pratiche e percezioni dei docenti universitari. *Research Trends in Humanities: Education & Philosophy*.
- Doria B., Picasso F. (2024). Alternative Assessment and Technology Enhanced Assessment practices: Research to inform Faculty Development processes. *QWERTY-Interdisciplinary Journal of Technology, Culture and Education*, 19(1): 52-71.
- Doria B., Grion V., and Paccagnella O. (2023). Pratiche valutative nelle università italiane: Una ricerca esplorativa a livello nazionale. *Italian Journal of Educational Research*, 30: 129-143.
- Doria B., Grion V., and Zaggia C. (2024). How to train university professors in assessment? A systematic literature review. *Form@re - Open Journal per la formazione in rete*, 24(2): 14-30.
- Facer K., Selwyn N. (2021). Digital technology and the futures of education – Towards ‘non-stupid’ optimism. *Educational Philosophy and Theory*, 53(3): 233-244.
- Florentino M.G., Montone A., and Ricciardiello G. (2024). The Feedback in a Formative Assessment Path: Development of Communicative Skills in a Workshop Online. In: *International Conference on Higher Education Learning Methodologies and Technologies Online* (pp. 250-261). Cham: Springer Nature Switzerland. DOI: 10.1007/978-3-031-67351-1.
- Giannandrea L. (2019). Valutazione, feedback, tecnologie. In: Rivoltella P.C. and Rossi P.G., a cura di, *Tecnologie per l'educazione* (pp. 69-81). Milano: Pearson.
- Giannandrea L., Ferrari S., and Laici C. (2024). Tecnologie per la riflessione e l'autovalutazione. E-portfolio, feedback, open badge. In: Rivoltella P.C. and Rossi P.G., a cura di, (nuova edizione), *Tecnologie per l'educazione* (pp. 189-199). Milano: Pearson.
- Grion V., Cesareni D. (2016). Multiplicity, fluidity, dialogue and sharing: Keywords to understand the complex dynamics between human learning and technology. *QWERTY-Interdisciplinary Journal of Technology, Culture and Education*, 11(1): 5-10.
- Grion V., Serbati A. (2019). *Valutazione sostenibile e feedback nei contesti universitari. Prospettive emergenti, ricerche e pratiche*. Lecce: PensaMultimedia.
- Grion V., Serbati A., Doria V., and Nicol D. (2021). Rethinking assessment and feedback practices in higher education: A review of recent literature. *Innovations in Education and Teaching International*, 58(4): 405-416.
- Gupta P., Ding B., Guan C., and Ding D. (2024). Generative AI: a systematic review using topic modelling techniques. *Data and Information Management*, 8(2), 100066. DOI: 10.1016/j.dim.2024.100066.
- Laici C. (2021). *Il feedback come pratica trasformativa nella didattica universitaria*. Milano: FrancoAngeli.
- Laici C., Pentucci M. (2023). Developing university students' feedback literacy through peer feedback activities. *Education Sciences & Society - Open Access*, 14(1). DOI: 10.3280/ess1-2023oa15925.
- Lipnevich A.A., Panadero E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, 6, 720195. DOI: 10.3389/feduc.2021.720195.

- Nicol D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35(5): 501-517.
- Nicol D. (2018). Unlocking generative feedback via peer reviewing. In: Grion V. and Serbati A., a cura di, *Assessment of learning or assessment for learning? Towards a culture of sustainable assessment in HE* (pp. 73-85). Lecce: Pensa MultiMedia.
- Nicol D. (2021). Guiding learning by activating students' inner feedback. *Times Higher Education*. <https://www.timeshighereducation.com/campus/guide-learning-activating-students-inner-feedback>.
- Nicol D.J., Macfarlane-Dick D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2): 199-218.
- Nieminen J.H., Carlless D. (2023). Feedback literacy: una revisione critica di un concetto emergente. *High Education*, 85: 1381-1400. DOI: 10.1007/s10734-022-00895-9.
- Panadero E., Lipnevich A.A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35, 100416.
- Panadero E., Fraile J., Fernández Ruiz J., Castilla-Estévez D., and Ruiz M.A. (2019). Spanish university assessment practices: Examination tradition with diversity by faculty. *Assessment & Evaluation in Higher Education*, 44(3): 379-397.
- Panciroli C., Rivoltella P.C. (2023). *Pedagogia algoritmica. Per una riflessione educativa sull'Intelligenza Artificiale*. Brescia: Scholè.
- Picasso F., Doria B., Grion V., Venuti P., and Serbati A. (2023). What technology-enhanced assessment and feedback practices do Italian academics declare in their syllabi? Analysis and reflections to support academic development. In: Fulantelli G., Burgos D., Casalino G., Cimitile M., Lo Bosco G., and Taibi D., a cura di, *Higher education learning methodologies and technologies online* (pp. 267-279). Cham: Springer.
- Price M., Handley K., Millar J., and O'Donovan B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, 35(3): 277-289.
- Raffaghelli J.E., Ghislandi P., Sancassani S., Canal L., Micciolo R., Balossi B., and Zani M. (2018). Integrating MOOCs in physics preliminary undergraduate education: beyond large size lectures. *Educational Media International*, 1-16. DOI: 10.1080/09523987.2018.1547544.
- Raffaghelli J.E. (2024). Post Digital Scholarship. Professionalità accademica e trasformazione digitale in università. *Didattiche, Tecnologie e Media Education. Frontiere per la Sostenibilità*, 1: 1-460.
- Ranieri M., Raffaghelli J.E., and Bruni I. (2021). Game-based student response system: Revisiting its potentials and criticalities in large-size classes. *Active Learning in Higher Education*, 22(2): 129-142. DOI: 10.1177/1469787418812667.
- Redecker C., Punie Y. (2017). *European framework for the digital competence of educators: DigCompEdu*. Lussemburgo: Publications Office of the European Union.

- Rossi P.G., Pentucci M., Fedeli L., Giannandrea L., and Pennazio V. (2018). Dal feedback informativo al feedback generativo. *Education Sciences and Society*, 9(2): 83-107.
- Sadler R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5): 535-550.
- Steiss J., Tate T., Graham S., Cruz J., Hebert M., Wang J., Moon Y., Tseng W., Warschauer M., and Olson C.B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
- Tam A.C.F. (2024). Interacting with ChatGPT for internal feedback and factors affecting feedback quality. *Assessment & Evaluation in Higher Education*, 1-17. DOI: 10.1080/02602938.2024.2374485.
- Tonelli D., Grion V., and Serbati A. (2018). L'efficace interazione fra valutazione e tecnologia: Evidenze da una rassegna sistematica della letteratura. *Italian Journal of Educational Technology*, 26(3): 6-23.
- Winstone N., Carless D. (2019). *Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach* (1st ed.). Routledge. DOI: 10.4324/9781351115940.