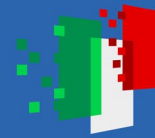




Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Future
Artificial
Intelligence
Research

Xai-driven knowledge distillation of Large Language Models for efficient deployment on low-resource devices

Riccardo Cantini, Alessio Orsino,
Domenico Talia

Università della Calabria

23-24 Settembre, Napoli



Introduzione

- I **Large Language Models** hanno riscosso notevole successo per le loro elevate capacità di comprensione e generazione del linguaggio naturale.
- Tuttavia essi richiedono elevate risorse computazionali, il che ostacola il loro utilizzo in contesti low-resource (e.g., edge AI).
- Tecniche di compressione:
 - **Pruning**: eliminazione di componenti superflue del modello.
 - **Quantization**: riduzione della precisione numerica dei pesi.
 - **Knowledge Distillation**: trasferimento della conoscenza da un modello di grandi dimensioni (i.e., *teacher*) ad uno più compatto ed efficiente (i.e., *student*).



Knowledge distillation

- La conoscenza viene trasferita minimizzando una loss che combina due contributi:
 - Task loss** (L_{CE}): misura l'errore del modello *student* rispetto ai dati annotati, ottimizzandone le prestazioni sul task specifico.
 - Distillation loss** (L_{KL}): misura la divergenza tra le predizioni del modello *student* e quelle del *teacher*, incoraggiando lo *student* ad imitarne il comportamento.

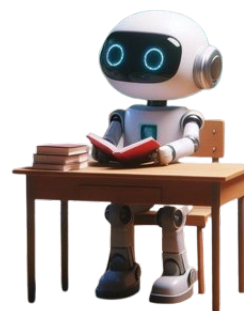
- Task: «Classifica una recensione come *positiva* o *negativa*». Classi: 0 (**negativa**), 1 (**positiva**).

R1: «Buon prodotto, fa il suo dovere.»

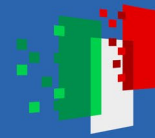
R2: «Prodotto eccellente, perfetto in ogni aspetto.»



$$L = (1 - \alpha)L_{CE} + \alpha L_{KL}^{\tau}$$



	R1	R2
Target label (WHAT)	1	1
Teacher output (HOW)	0.75	0.99



Problemi degli approcci attuali

- Le tecniche di distillazione tradizionali non sono sempre in grado di trasferire in maniera efficace la «*conoscenza explainable*» da modelli complessi (e.g., LLMs) a modelli leggeri.
- Il semplice allineamento degli output dei due modelli potrebbe fallire nel trasferire allo *student* informazioni essenziali sul processo decisionale in accordo al quale il *teacher* svolge uno specifico task.
 - Perdita di **interpretabilità**
 - Basse capacità di **generalizzazione**
- Soluzione proposta: **DiXtill** (*XAI-driven Knowledge Distillation*)
 - Le spiegazioni locali di un LLM *teacher* vengono utilizzate per guidare il processo di distillazione in un modello *student* **energy-efficient** e **self-explainable**.
 - Questo approccio migliora la **trustworthiness** dello *student*, con un conseguente impatto positivo sull'**accuratezza** del modello distillato sul task specifico.

DiXtill: XAI-driven knowledge distillation

- Alla tradizionale loss di distillazione viene aggiunto un termine (L_{XAI}) che promuove l'allineamento tra le spiegazioni *locali* del *teacher* e dello *student*.
 - Le spiegazioni del modello *teacher* (σ^T) sono precalcolate mediante l'uso di una tecnica di XAI post-hoc.
 - Le spiegazioni del modello *student* (σ^S) vengono apprese in maniera dinamica durante il processo stesso di distillazione.

- Task: «Classifica una recensione come *positiva* o *negativa*». Classi: 0 (**negativa**), 1 (**positiva**).

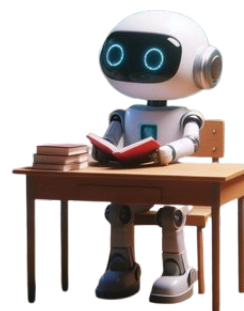
R1: «Buon prodotto, fa il suo dovere.»

R2: «Prodotto eccellente, perfetto in ogni aspetto.»



$$L = (1 - \alpha)L_{CE} + \alpha(L_{KL}^T + L_{XAI})$$

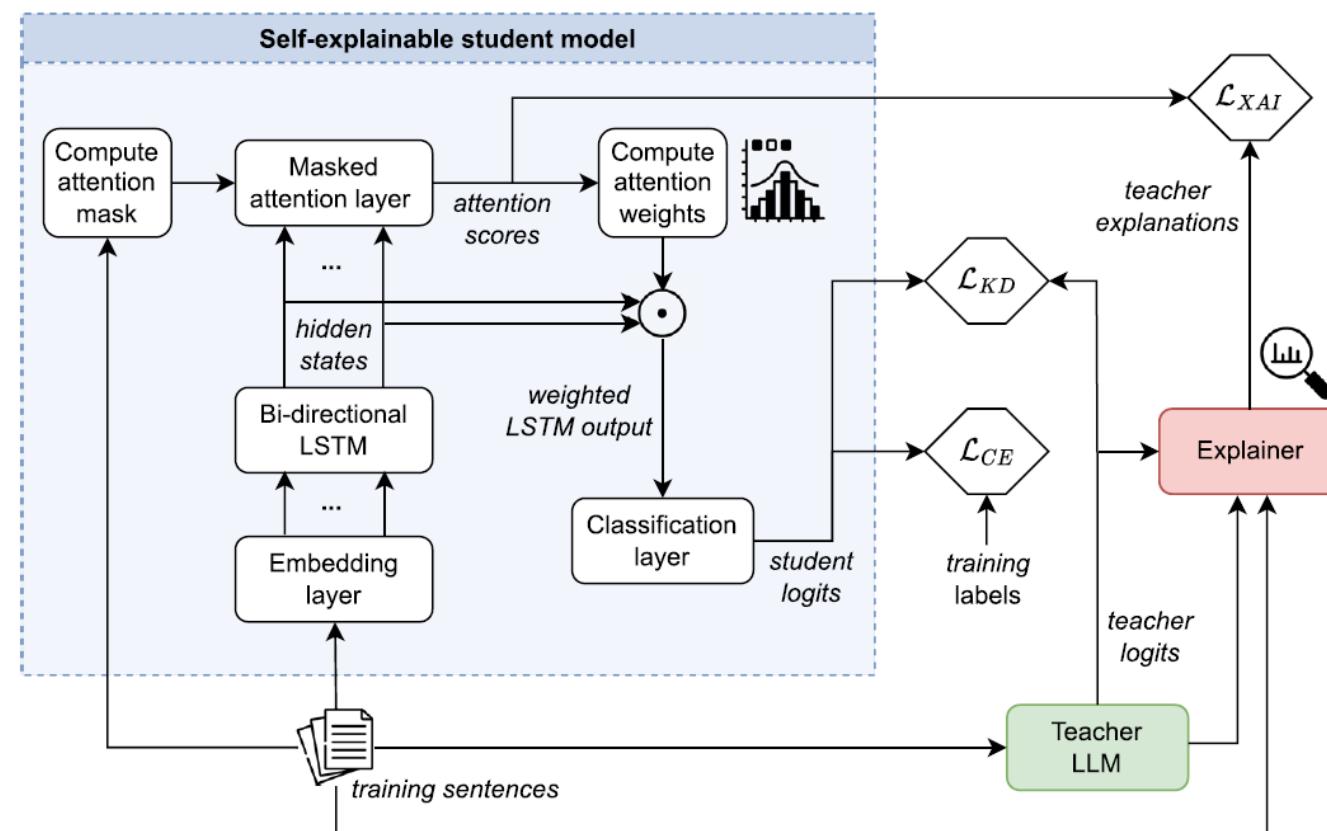
$$L_{XAI} = \frac{1}{2} \left(1 - \frac{\sigma^T \cdot \sigma^S}{\|\sigma^T\| \|\sigma^S\|} \right)$$



	R1	R2
Target label (WHAT)	1	1
Teacher output (HOW)	0.75	0.99
Teacher expl. (WHY)	<i>buon</i>	<i>eccellente</i>

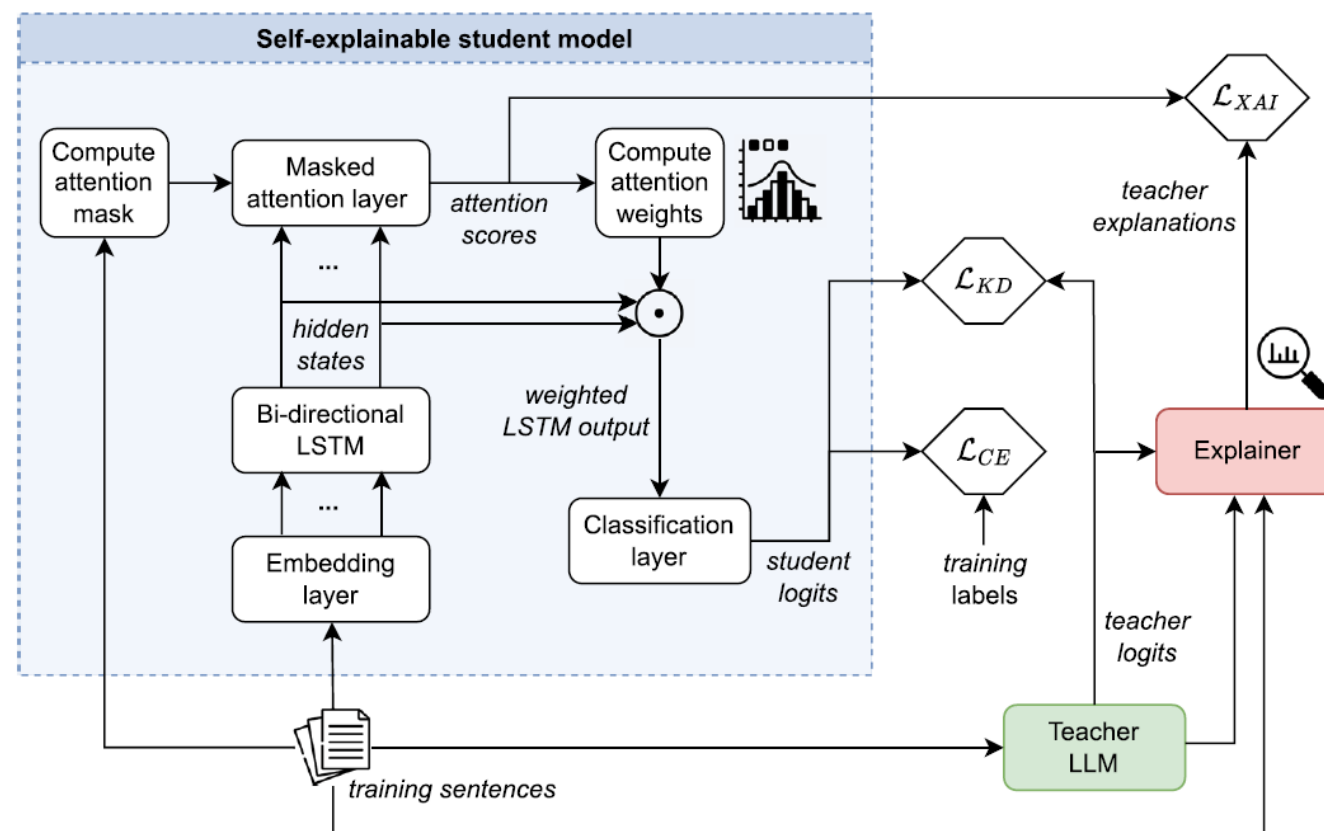
DiXtill: struttura del processo di distillazione

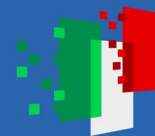
- Modello *student*: LSTM bi-direzionale
 - Il modello *student* è **self-explainable**, ovvero, dato un certo input, fornisce in output sia un risultato sia la relativa spiegazione.
 - Sfrutta un meccanismo di **masked attention** per attribuire ad ogni termine uno *score* che misura la sua importanza nel determinare l'output.
 - Tali score (**word attributions**) costituiscono una spiegazione dell'output del modello *student*.
 - Il **masking** consente allo *student* di ignorare elementi superflui, come il padding, durante il calcolo degli score di attention.



DiXtill: struttura del processo di distillazione

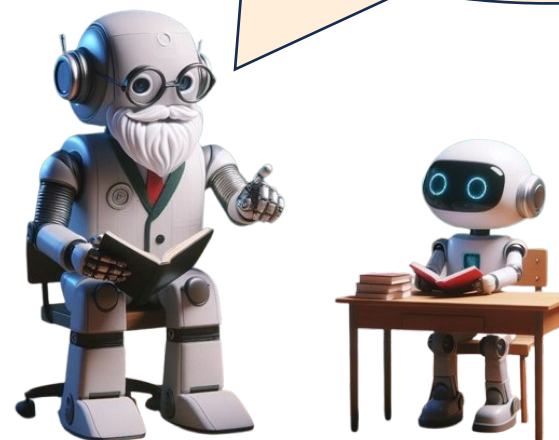
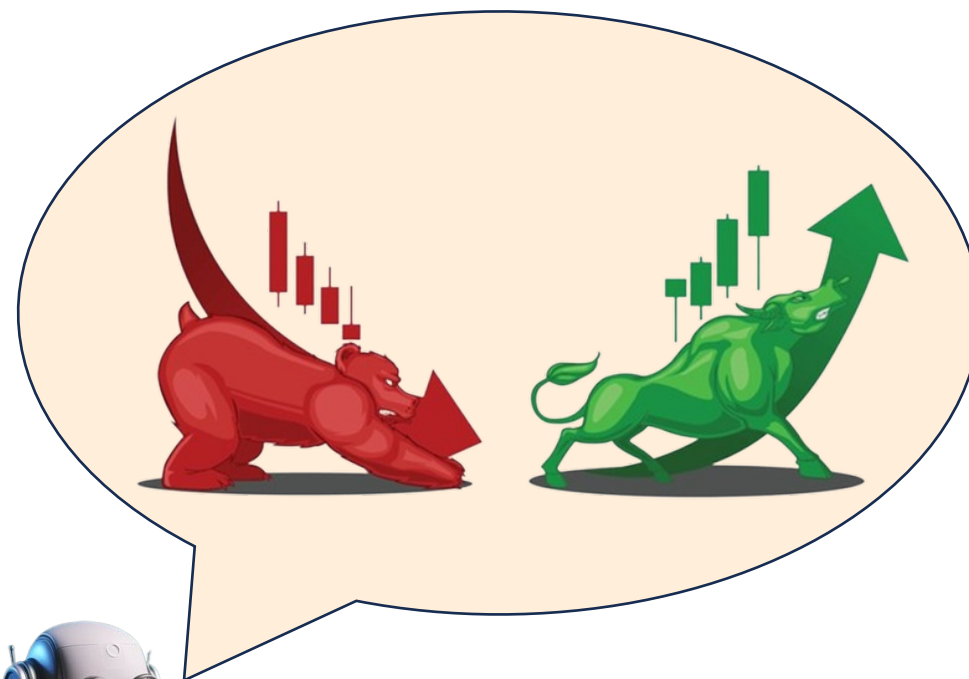
- Modello *teacher*: BERT
 - Il modello *teacher* è **preaddestrato** e **fine-tuned** sul task di interesse (e.g., *sentiment analysis*).
 - Viene usato un **Explainer** (e.g., *integrated gradients*) per precalcolare le spiegazioni del *teacher* per ogni elemento del training set.
 - Le spiegazioni ottenute, sotto forma di **word attributions**, vengono utilizzate per guidare il processo di distillazione (L_{XAI}).
- DiXtill loss: $L = (1 - \alpha)L_{CE} + \alpha(L_{KL}^\tau + L_{XAI})$





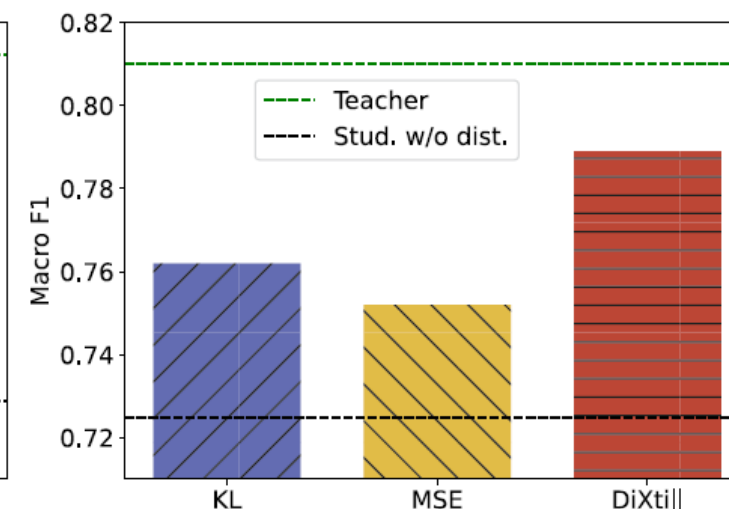
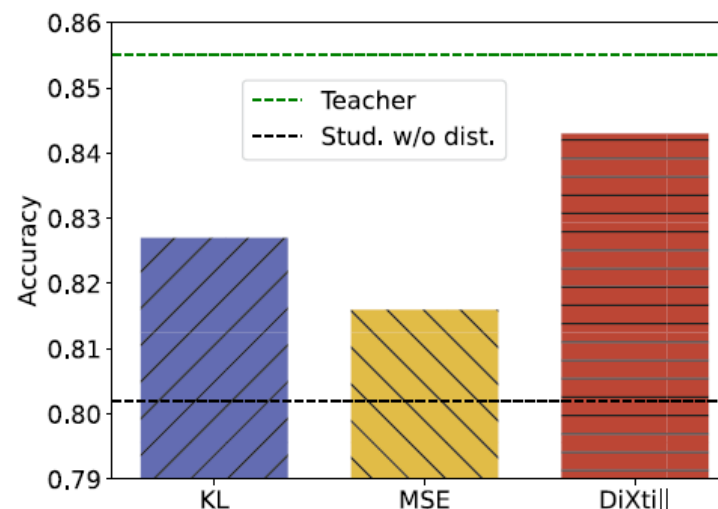
Valutazione sperimentale

- Dataset utilizzato: **Twitter Financial News Sentiment**
 - Classi: **bearish** (*ribassista*), **bullish** (*rialzista*), **neutral**
 - 9.938 tweet di training, 2.486 tweet di test
 - Modello fine-tuned: **FinBERT** (🧡 **Hugging Face**)
- Tecniche confrontate:
 - **Distillazione logit-based** (KL / MSE loss)
 - **Post-Training Quantization** (PTQ), int8
 - **Attention Head Pruning** (AHP), structured



Confronto con le tecniche di distillazione classica

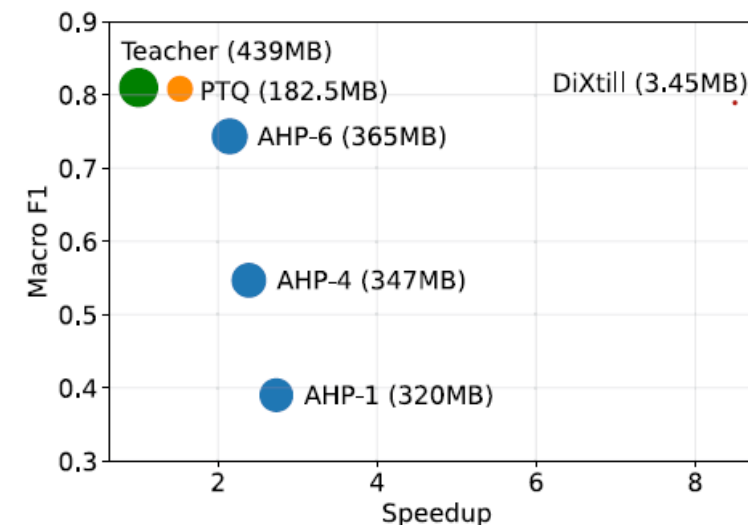
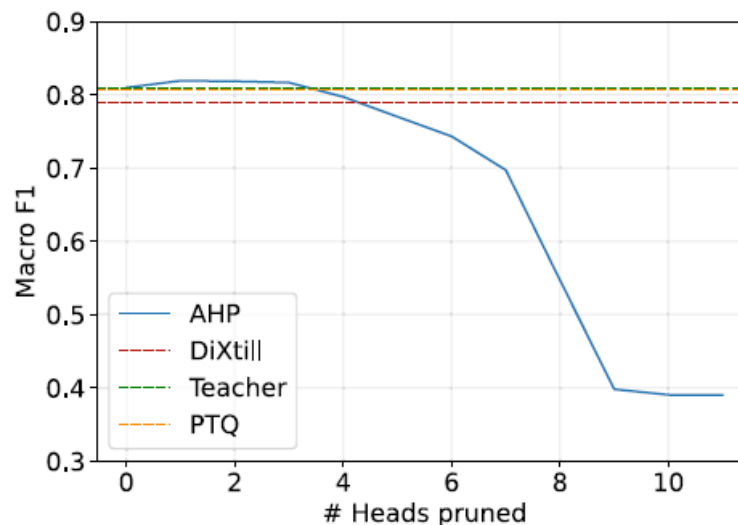
- **DiXtill** presenta le prestazioni migliori, con un'accuratezza di **0.843** e una **macro-F1** di **0.789**.
- L'integrazione delle spiegazioni nel processo di distillazione consente di ridurre il divario prestazionale tra *teacher* e *student*.
- Vengono mantenute prestazioni elevate a fronte di una importante riduzione del numero di parametri (*meno di un milione contro i 110 milioni del modello teacher*).



Method	Accuracy	Macro F1
Student w/o distillation	0.802	0.725
Distillation with KL	0.827	0.762
Distillation with MSE	0.816	0.752
DiXtill	0.843	0.789
Teacher	0.855	0.810

Confronto con le tecniche di compressione

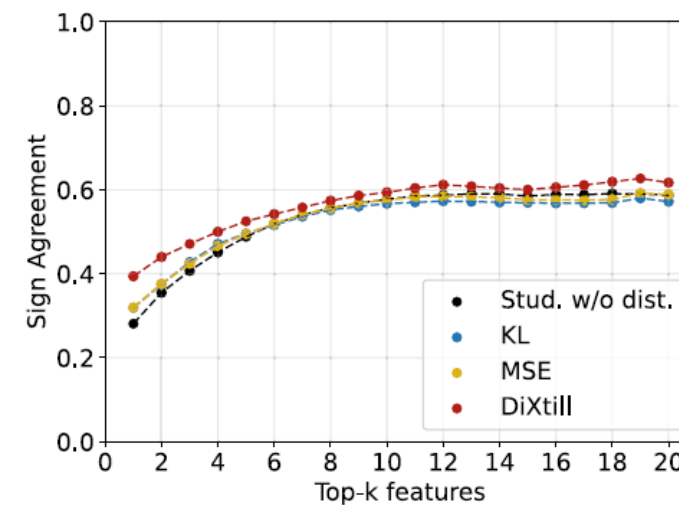
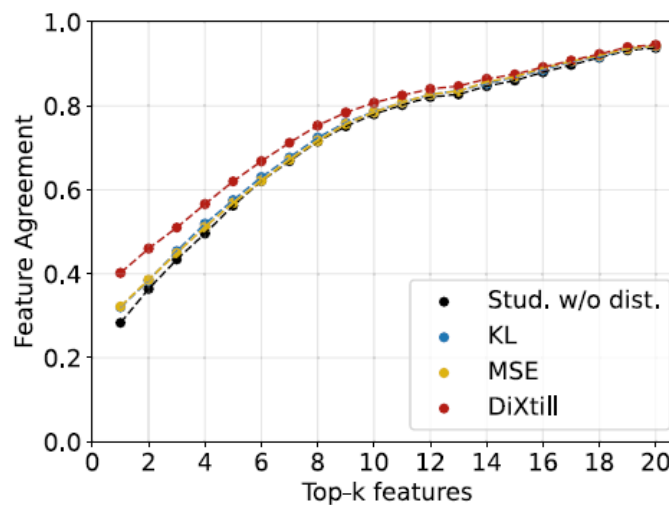
- **DiXtill** raggiunge il miglior trade-off tra accuratezza ed efficienza in termini di **compressione (127x)** e **tempo di inferenza (8.7x)**.
- La quantizzazione **PTQ** ottiene una buona accuratezza, al prezzo di uno speed-up (1.52x) ed una compressione limitati (2.4x).
- Il pruning **AHP** migliora la velocità rispetto a PTQ (2.18x), ma risulta molto sensibile al numero di *heads* rimosse, con prestazioni insufficienti per elevati fattori di compressione.



Method	Size (C_{ratio})	Inference time (<i>Speedup</i>)
AHP-6	365 MB (↑ 1.20x)	0.28 s (↑ 2.18x)
PTQ	182.5 MB (↑ 2.40x)	0.40 s (↑ 1.52x)
DiXtill	3.45 MB (↑ 127x)	0.07 s (↑ 8.7x)
Teacher	439 MB	0.61 s

Interpretabilità del modello distillato

- **DiXtill** ha mostrato un maggiore accordo (*feature e sign agreement*) tra le spiegazioni dei modelli *student* e *teacher*, rispetto alle altre tecniche di distillazione.
- Le spiegazioni per **DiXtill** sono *self-computed*, mentre per le altre tecniche sono ottenute tramite *IG*.



Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (1.00)	BEARISH	2.68	nortonlifelock stock price target cut to \$ 18 from \$ 25 at deutsche bank

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (1.00)	BULLISH	2.46	autodesk stock price target raised to \$ 162 from \$ 149 at wedbush

- Esempi di spiegazioni:
 - Tweet con sentiment ribassista: "cut", "stock", "price".
 - Tweet con sentiment rialzista: "raised", "stock", "price".

Conclusioni e sviluppi futuri

- **DiXtill** consente la distillazione efficace di LLM in modelli di piccole dimensioni seguendo un approccio XAI-driven.
- Vantaggi:
 - Elevata **accuratezza** del modello distillato e maggiore accordo con le spiegazioni del teacher rispetto alla distillazione classica.
 - Fattore di **compressione** e **speed-up** significativamente più alti rispetto ad altre tecniche di compressione (*PTQ, AHP*).
- Sviluppi correnti e futuri:
 - Integrazione con tecniche di **meta-learning** (*learning-to-teach*).
 - Combinazione con tecniche di **neural architecture search** green-aware.
- Codice disponibile su GitHub: <https://github.com/SCALabUnical/DiXtill>
- Maggiori dettagli in: *Cantini Riccardo, Alessio Orsino, and Domenico Talia. "Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices." Journal of Big Data 11.1 (2024): 63.*

