clearbox AI

Turin, 2023

Come rilevare automaticamente testo generato da Large Language Models

humans && machines

Introduzione

Clearbox AI e' un fornitore di servizi per la generazione di dati sintetici.

Dati sintetici: dati finti ma realistici molto spesso creati tramite modelli generativi.

Un modo per valutare **qualità dati sintetici** e' tramite l'allenamento di algoritmi di **detection**.

Simona Mazzarino, collega impegnata sul fronte generazione dati testuali.

Importanza algoritmi di detection in vari contesti

ESEMPI

- **Ambito accademico**: articoli scientifici potrebbero includere intere sezioni generate tramite LLM. Difficoltà nel reperire peer-reviewers aumenta il rischio che articoli pubblicati contengano nozioni sbagliate o riferimenti inesistenti.

Importanza algoritmi di detection in vari contesti

ESEMPI

- **Ambito accademico**: articoli scientifici potrebbero includere intere sezioni generate tramite LLM. Difficoltà nel reperire peer-reviewers aumenta il rischio che articoli pubblicati contengano nozioni sbagliate o riferimenti inesistenti.

- **Contenuti online**: ad esempio recensioni finte. Come facciamo ad evitare che siti/piattaforme contenenti recensioni utenti non vengano inondati ulteriormente da contenuti generati da bot?

Importanza algoritmi di detection in vari contesti

ESEMPI

- **Ambito accademico**: articoli scientifici potrebbero includere intere sezioni generate tramite LLM. Difficoltà nel reperire peer-reviewers aumenta il rischio che articoli pubblicati contengano nozioni sbagliate o riferimenti inesistenti.

- **Contenuti online**: ad esempio recensioni finte. Come facciamo ad evitare che siti/piattaforme contenenti recensioni utenti non vengano inondati ulteriormente da contenuti generati da bot?

- **Scrittura proposal**: come valutare la presenza di testo generato artificialmente all'interno di proposte che partecipano a gare per finanziamenti?

Possibili approcci per rilevare testo generato artificialmente

Tre esempi, presi dalla newsletter ***The Batch (deeplearning.ai)***:

- Watermarking
- Analisi probabilita' output language models
- Algoritmi di classificazione

Watermarking

Idea: fornitori di servizi di generazione basati su Large Language Models possono includere un'impronta invisibile all'interno del testo generato.

A Watermark for Large Language Models

John Kirchenbauer * Jonas Geiping * Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

https://arxiv.org/abs/2301.10226

Watermarking

Metodo:

1. Token associato a ultima parola generata viene usato per creare il seed per un generatore di numeri random.

2. Generatore random usato per assegnare al 20% delle parole del vocabolario da cui campionare il token successivo una probabilita' di campionamento minore, insieme di queste parole viene inserito in una blacklist.

3. Analisi delle occorrenze delle parole incluse in queste blacklist all'interno di un testo definisce probabilita' che il testo sia generato

Watermarking

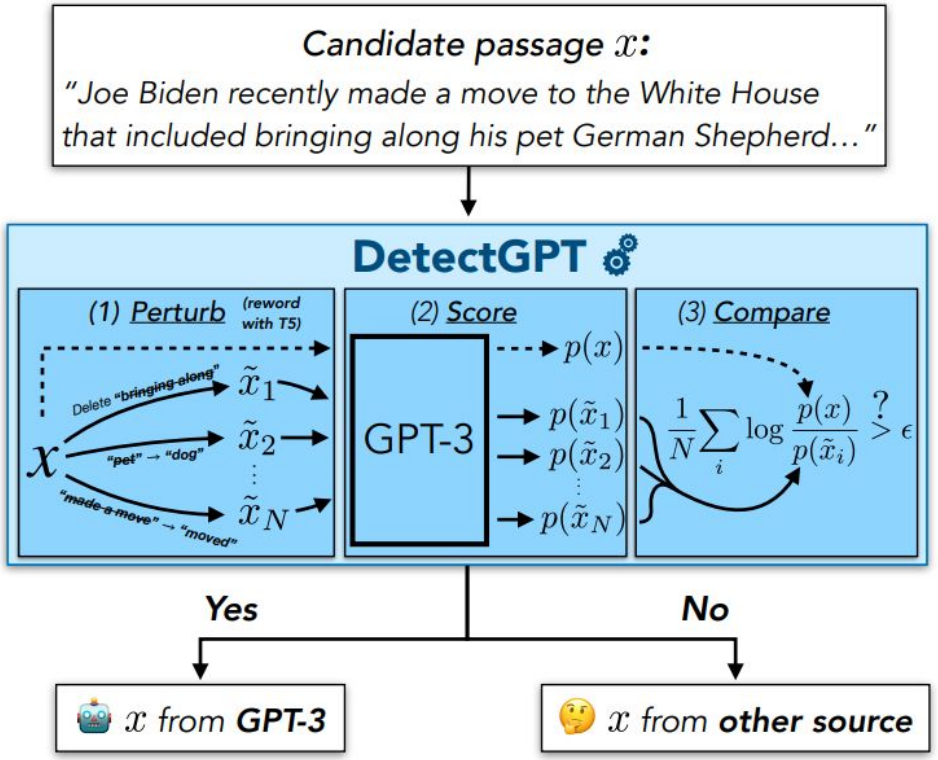
Vantaggi-Svantaggi:

- Watermark deve essere implementato lato fornitore LLM
- Permette di calcolare probabilita' che singoli estratti di testo contengano watermark
- Piccola riduzione in qualità del testo generato

Likelihood analysis

Metodo:

1. Generazione di N perturbazioni intorno a una frase da analizzare (usando LLM)
2. Calcolo delle likelihoods associate a diverse perturbazioni secondo un LLM
3. Stima della probabilitá che il testo provenga da LLM in base a likelihoods collezionate

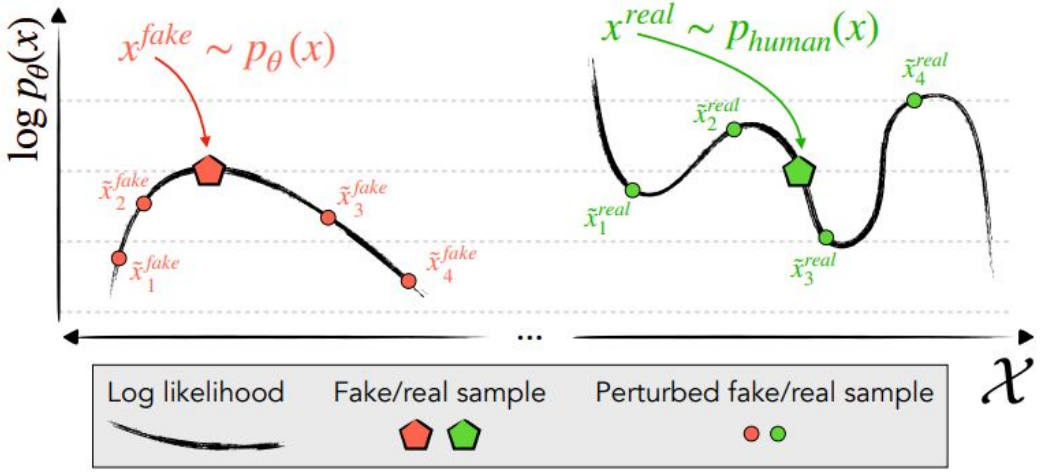


https://arxiv.org/abs/2301.11305

Likelihood analysis

Vantaggi-svantaggi

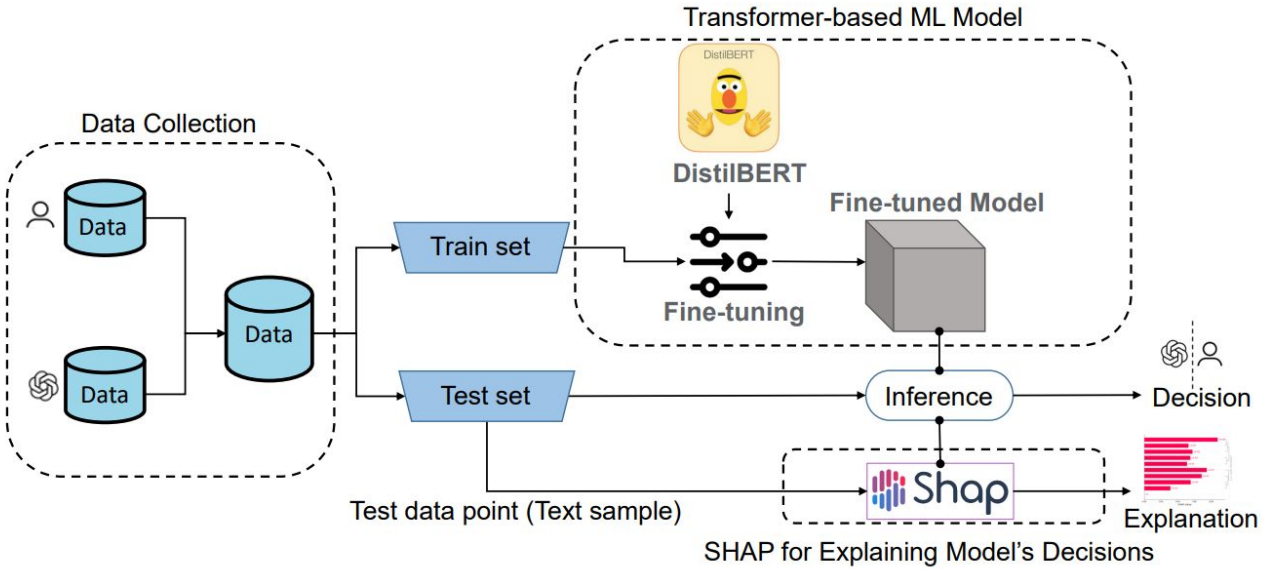
1. Computazionalmente pesante (generazione perturbazioni e calcolo likelihoods)
2. Per ora testato solo su GPT-2?



https://arxiv.org/abs/2301.11305

Classificazione

Idea: allenare un **discriminatore ML** a distinguere tra testo umano e testo generato.



https://arxiv.org/abs/2301.13852

Classificazione

Idea: allenare un **discriminatore ML** a distinguere tra testo umano e testo generato.

Classificatore deve operare in un **contesto preciso** (ad esempio rilevazione recensioni finte o contenuti accademici).

Affetto da stesse problematiche che caratterizzano classificatori ML

- Spiegabilita’
- Calibrazione
- Data drift

Parentesi:
Analisi interpretativa e calibrazione

Analisi interpretativa: Shapley values

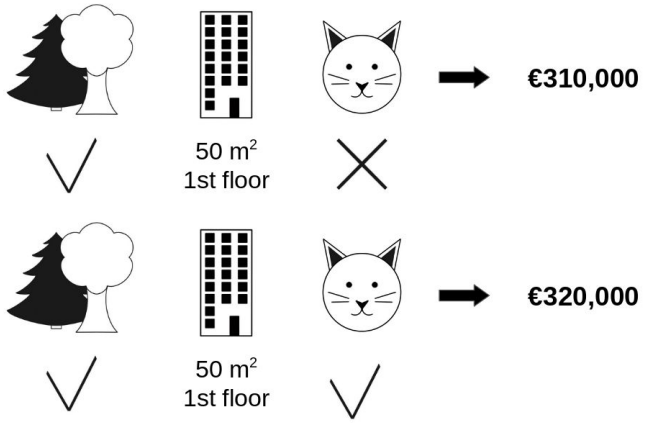
Concetto proveniente dalla teoria dei giochi: *come ridistribuire la ricompensa di un gioco a cui ha partecipato un gruppo di giocatori in maniera cooperativa*?

Coefficienti di Shapley definiscono una maniera per distribuire ricompensa tra partecipanti.

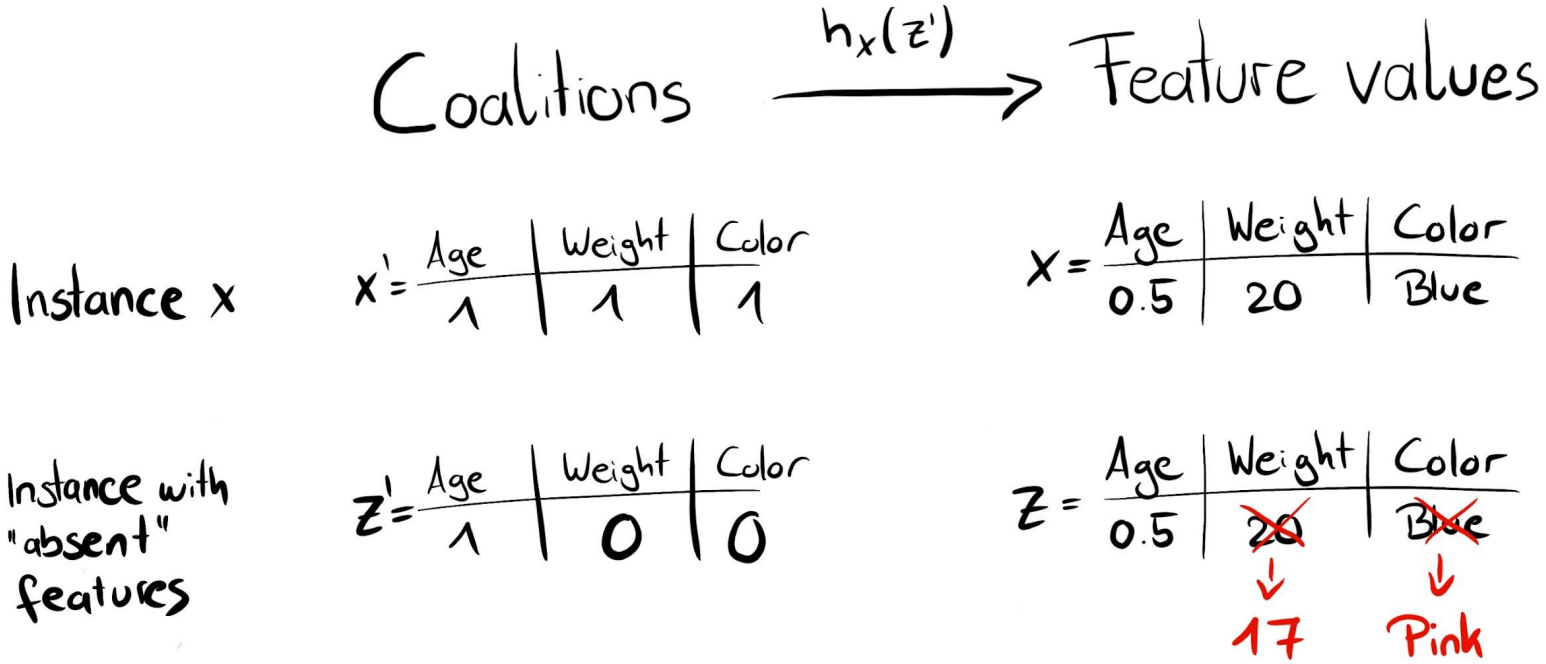
Applicato a language models:

Giocatori → Parole o tokens

Ricompensa → Output del modello



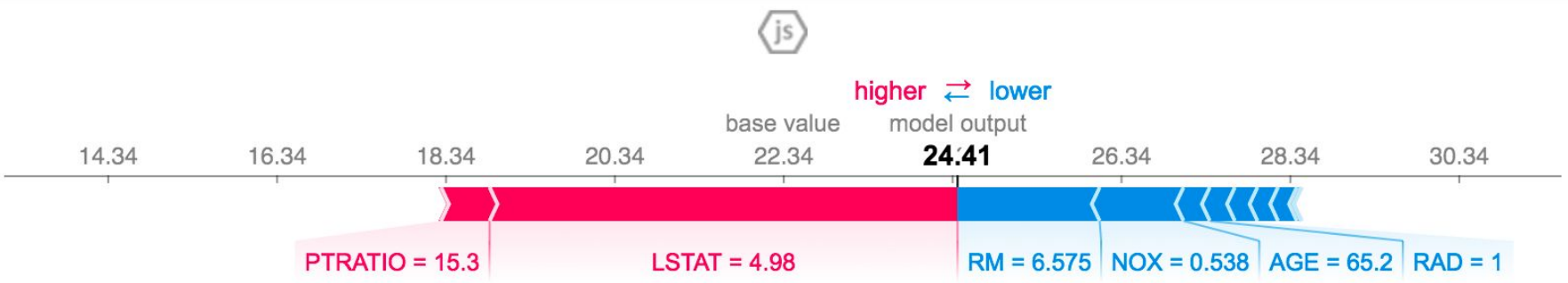
https://christophm.github.io/interpretable-ml-book/



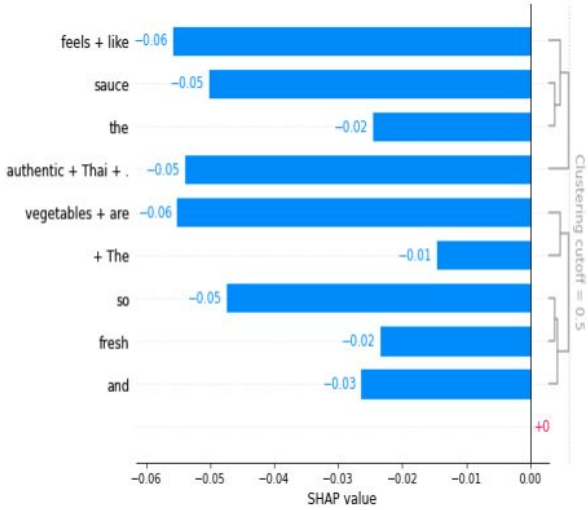
https://christophm.github.io/interpretable-ml-book/

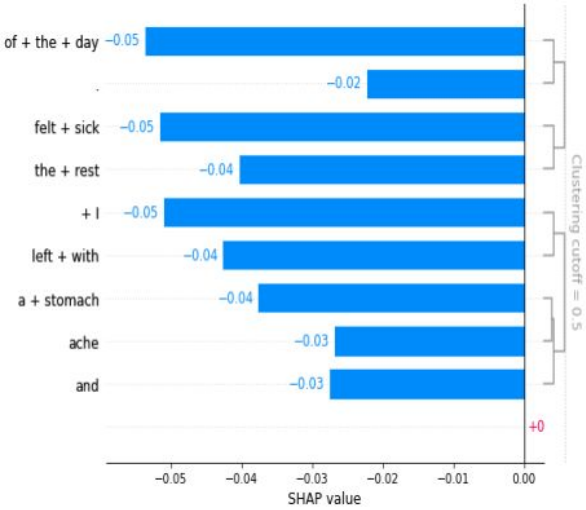
SHAP

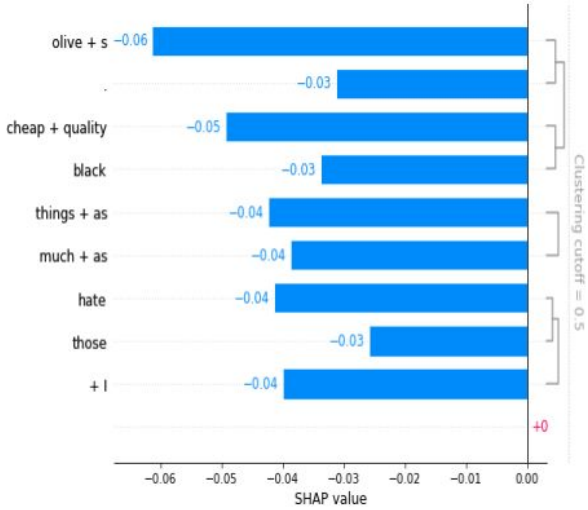
Approssimazione Shapley Values

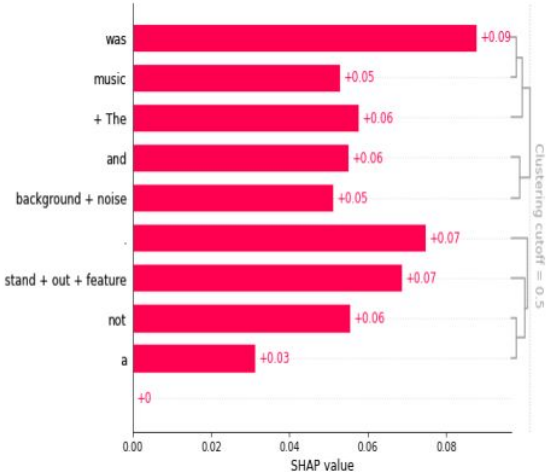


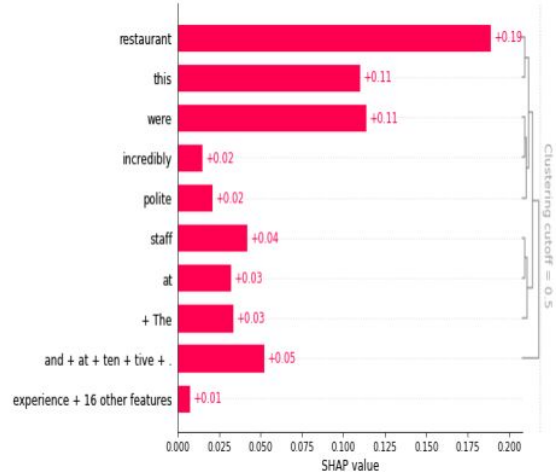
- Libreria SHAP offre oltre al metodo stesso librerie di visualizzazione molto curate
- KernelSHAP e' decisamente lento.

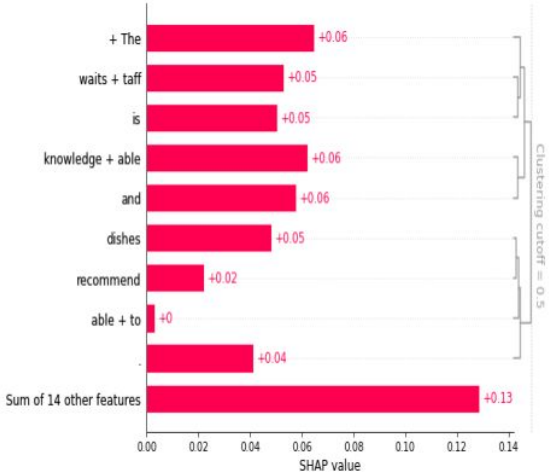

feels + like −0.06
sauce −0.05
the −0.02
authentic + Thai + . −0.05
vegetables + are −0.06
+ The −0.01
so −0.05
fresh −0.02
and −0.03
+0
−0.06 −0.05 −0.04 −0.03 −0.02 −0.01 0.00
SHAP value
Clustering cutoff = 0.5


of + the + day −0.05
. −0.02
felt + sick −0.05
the + rest −0.04
+ I −0.05
left + with −0.04
a + stomach −0.04
ache −0.03
and −0.03
+0
−0.05 −0.04 −0.03 −0.02 −0.01 0.00
SHAP value
Clustering cutoff = 0.5


olive + s −0.06
. −0.03
cheap + quality −0.05
black −0.03
things + as −0.04
much + as −0.04
hate −0.04
those −0.03
+ I −0.04
+0
−0.06 −0.05 −0.04 −0.03 −0.02 −0.01 0.00
SHAP value
Clustering cutoff = 0.5


was +0.09
music +0.05
+ The +0.06
and +0.06
background + noise +0.05
. +0.07
stand + out + feature +0.07
not +0.06
a +0.03
+0
0.00 0.02 0.04 0.06 0.08
SHAP value
Clustering cutoff = 0.5


restaurant +0.19
this +0.11
were +0.11
incredibly +0.02
polite +0.02
staff +0.04
at +0.03
+ The +0.03
and + at + ten + tive + . +0.05
experience + 16 other features +0.01
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
SHAP value
Clustering cutoff = 0.5


+ The +0.06
waits + taff +0.05
is +0.05
knowledge + able +0.06
and +0.06
dishes +0.05
recommend +0.02
able + to +0
. +0.04
Sum of 14 other features +0.13
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
SHAP value
Clustering cutoff = 0.5

Calibrazione modelli

Definizione

Approccio comune in ambito machine learning: incorporare le varie sorgenti di incertezza all'interno di un'unica quantità, la **confidenza del modello**.

Problema sorge quando in presenza di modelli cosiddetti non calibrati:

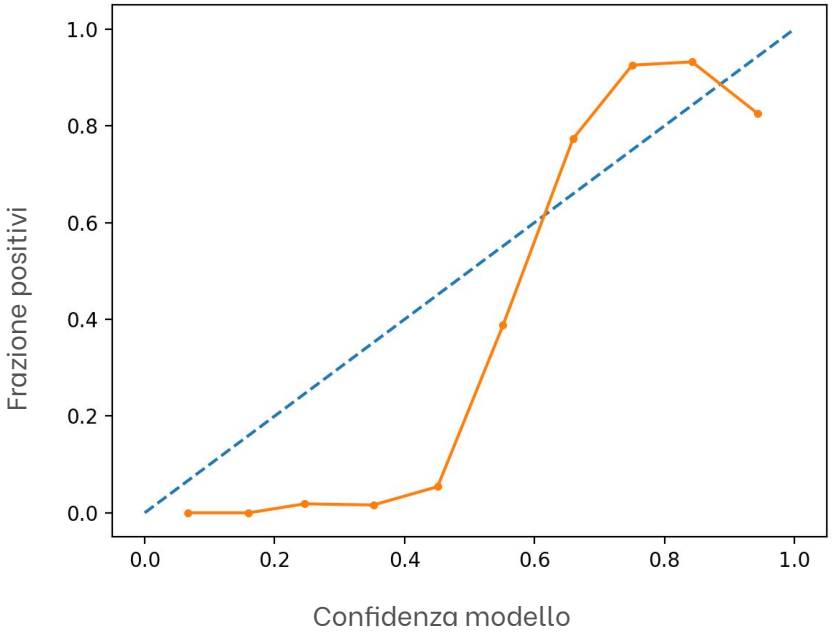
es. *una confidenza del 90% in un'etichetta deve corrispondere dal vivo in un rateo di successo analogo.*

Calibrazione modelli

Metodi

Modello di detection sarà molto probabilmente operato da un umano. E' perciò importante che il modello fornisca una **confidenza nel proprio output affidabile**.

Confidenza del modello molto importante in un contesto in cui l'output del modello può avere ripercussioni importanti sulle persone (ad esempio studenti accusati di usare chatGPT).

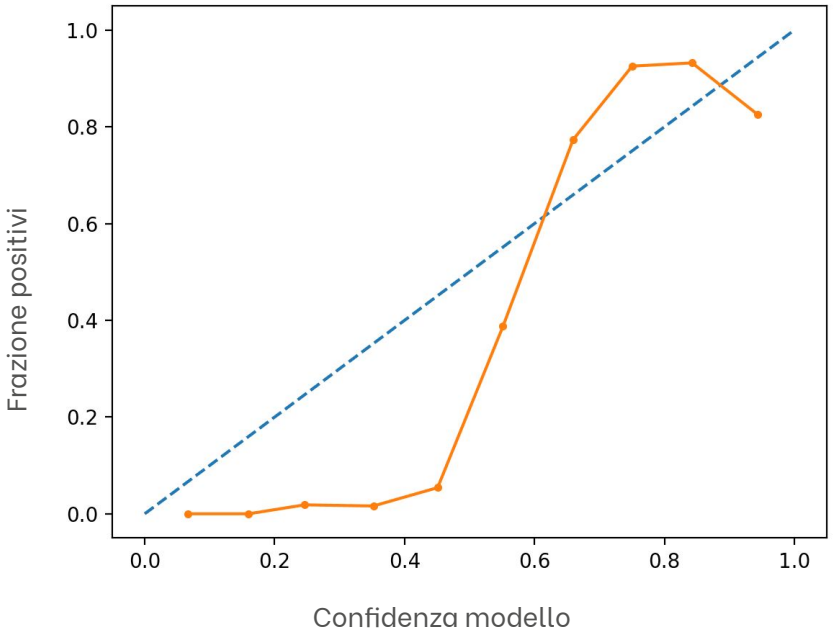


Calibrazione modelli

Metodi

Un modello che e' confidente al 90% che un testo sia stato generato da LLMs deve essere corretto 9 volte su 10 una volta utilizzato sul campo.

Calibrazione di un modello può essere quantificata tramite la curva di calibrazione (o reliability diagram)



Preparazione dataset

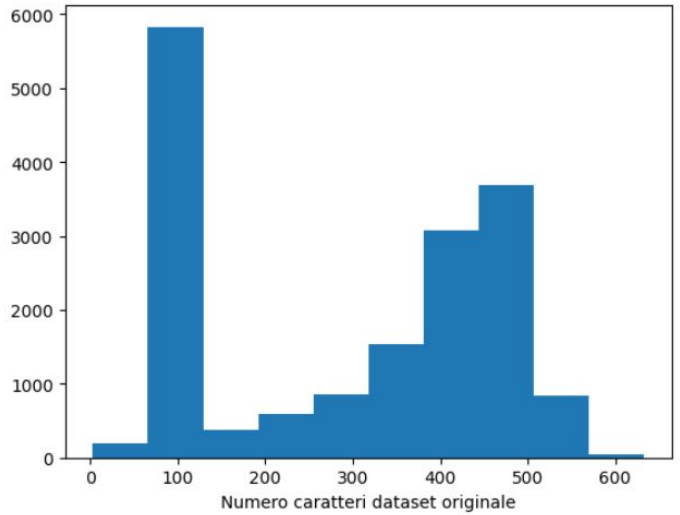
Il dataset di allenamento deve presentare il più fedelmente possibile il contesto in cui andra' a lavorare il classificatore.

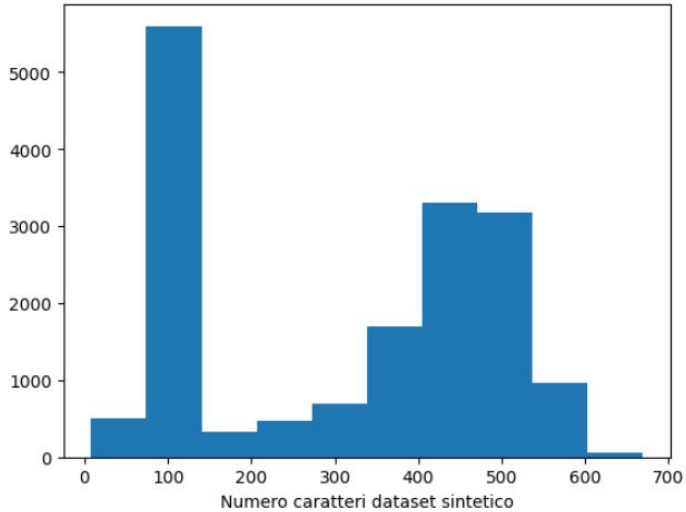
E' importante quindi che sia preparando tenendo conto della modalità con cui il testo generato che dobbiamo cercare di rilevare sarà ottenuto.

Preparazione dataset

Il dataset di allenamento deve presentare il più fedelmente possibile il contesto in cui andra' a lavorare il classificatore.

E' importante quindi che sia preparando tenendo conto della modalità con cui il testo generato che dobbiamo cercare di rilevare sarà ottenuto.





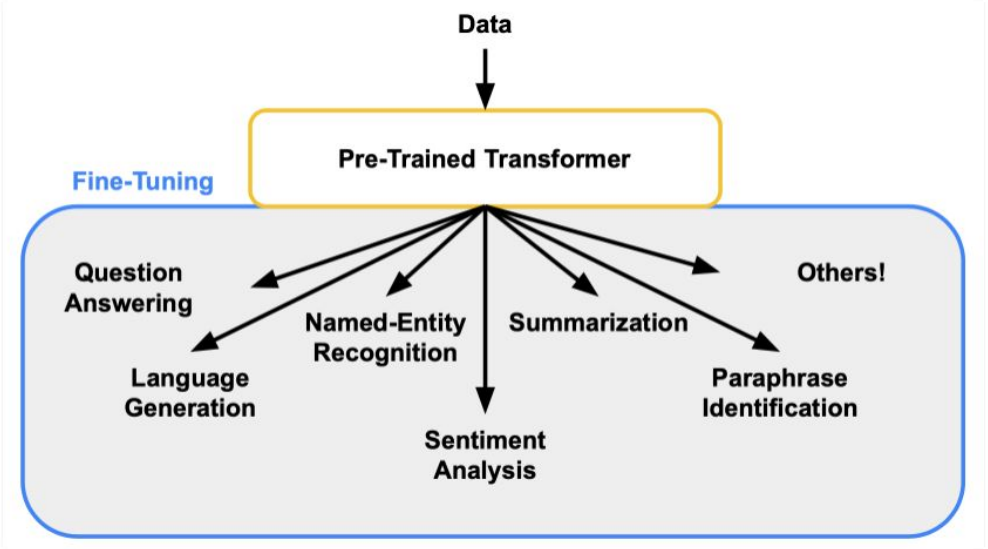
Esempi pratici

Esempio applicato a un dataset di detection proveniente dal mondo accademico.

1. Fine-tuning di un modello OpenAI
2. Allenamento di un semplice classificatore interpretabile
3. Fine-tuning di un classificatore basato su DistilBERT

Esempio: fine-tuning di un classificatore

Modelli LLMs già allenati possono essere utilizzati per risolvere task specifici usando un numero relativamente ridotto di esempi aggiuntivi.



https://www.assemblyai.com/blog/fine-tuning-transformers-for-nlp/

Esempi pratici

Cosa non siamo riusciti a coprire oggi (to be continued?)

Argomenti che abbiamo dovuto lasciare fuori:

- Come generare dataset di allenamento in maniera programmatica
- Fine-tuning di modelli più grandi (come GPT-J) tramite algoritmo **LoRa** (Low Rank Adaptation)
-

clearbox AI
humans && machines

Thanks for Reading

Feel free to contact us:


www.clearbox.ai


support@clearbox.ai


@ClearboxAI