Tagging automatico delle pagine web dei Comuni

7 Febbraio 2025

Dr. Davide Riva

Dipartimento di Informatica
Università degli Studi di Milano

Dr. Marzio De Corato

Dipartimento di Informatica
Università degli Studi di Milano

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
MUSA

Contenuti

1. Presentazione pipeline di lavoro
2. Risorse necessarie
3. Dati raccolti
4. Risultati ottenuti

Parte 1
Presentazione pipeline di lavoro

Principali problemi da affrontare

1. Scarsità di dati annotati (siti web con tag aggiunti manualmente)
2. Persistente eterogeneità nella struttura dei siti web (codice HTML)
3. Necessità di pulizia dei dati testuali
4. Similarità/ambiguità tra tag diversi

Pipeline iniziale

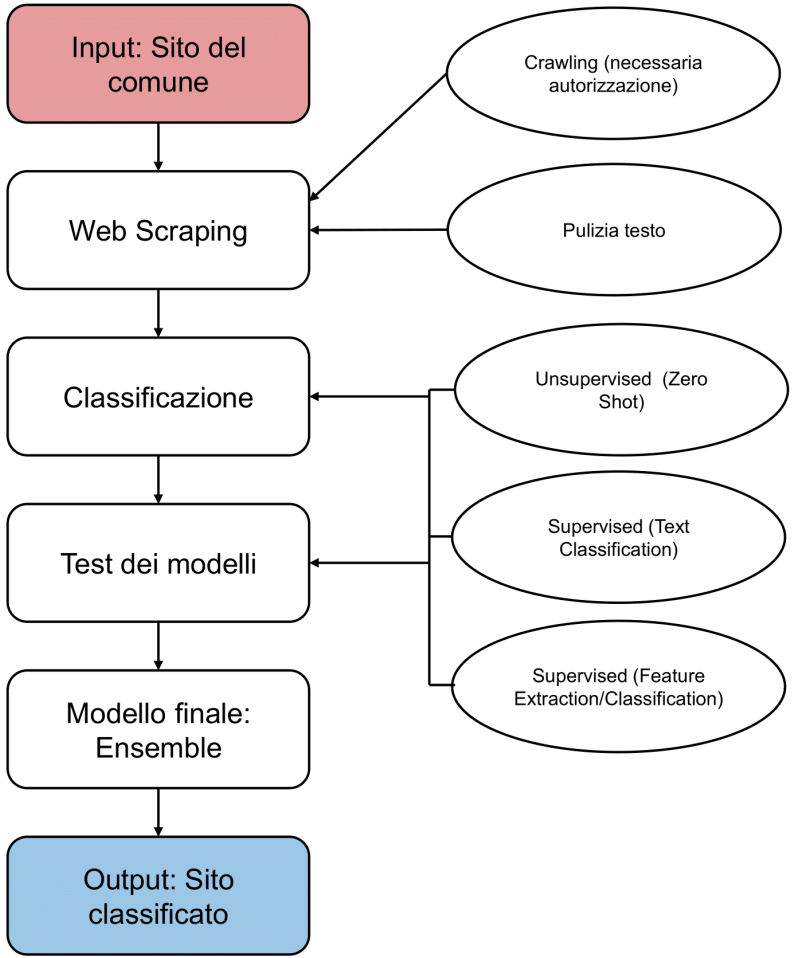
Ensemble di 3 modelli, la cui base comune è l'utilizzo di un Large Language Model (LLM) a cui sottoporre il testo contenuto nella specifica pagina web.

A. Zero-Shot
   → *A quale/i categoria/e appartiene il testo?*

B. Data Augmentation
   → *Genera altri testi simili*

C. Feature Extraction
   → Quali informazioni sono contenute nel testo?

L'approccio (A) è interamente non-supervisionato, mentre gli approcci (B) e (C) preparano l'input per un modello supervisionato (cioè addestrato su una parte dei dati e applicato ai dati restanti).

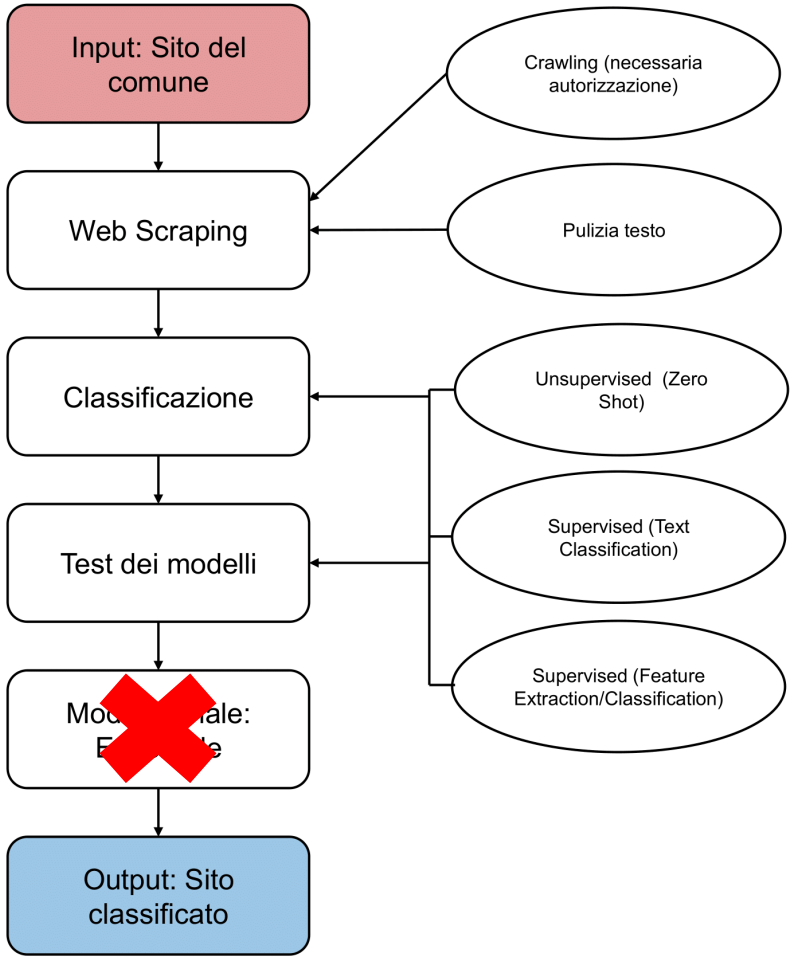


Pipeline iniziale

Sulla base dei risultati ottenuti, data l'elevata variabilità degli approcci (B) e (C), basati su Data Augmentation e Feature Extraction, l'idea di un modello ensemble è stata scartata, concentrandosi invece sul miglioramento ed efficientamento dei 3 approcci presi singolarmente.

In particolare:

- Pulizia dei testi estratti dalle pagine web (problema 3)
- Prompt engineering (problema 1)
- Post-processing dei risultati (problema 4)



I diritti dell’uomo: Cos’è persona? Cosa sono i diritti? – Comune di Gorgonzola Vai ai contenuti Vai al footer Regione Lombardia Accedi Comune di Gorgonzola Seguici su Youtube Youtube Cerca Cerca Apri burger menu Nascondi la navigazione Chiudi burger menu Comune di Gorgonzola Amministrazione Novità Servizi Vivere il Comune Tutti gli argomenti... Home / Vivere il Comune / Eventi / Evento di formazione / I diritti dell'uomo: Cos'è persona? Cosa sono i diritti? I diritti dell'uomo: Cos'è persona? Cosa sono i diritti? 24 ottobre Primo incontro di un percorso per giovani attraverso esperienze e testimonianze Condividi Facebook Twitter Linkedin Whatsapp Telegram Vedi azioni Stampa Ascolta Invia Vai al calendario eventi Indice della pagina Cos'è A chi è rivolto Luogo Date e orari Costi Allegati Contatti Ulteriori informazioni Cos'è Con l’educatore ____ e don ____ ci porremo alcune domande fondamentali all’inizio di questo percorso sui diritti dell’uomo per un confronto e un dibattiro di sviluppi futuri. A chi è rivolto A tutti a ragazzi e ragazze tra i 17 e i 18 anni in su. Luogo Centro Intergenerazionale Via Italia, 84 Gorgonzola MI, 20064 + – Leaflet Date e orari 24 Ott Inizio ore 21:00 Costi Partecipazione gratuita Ingresso gratuito Allegati LOCANDINA I diritti dell'uomo Contatti …

Pulizia dei testi estratti

Cosa eliminare?

- Elementi generali del sito (es: appartenenti all’header, al footer o a menù interni alla pagina)
- Rimandi al Comune (eccezione per gli indirizzi, significativi per indicare ad esempio un evento)
- Spazi vuoti e punteggiatura in eccesso

I diritti dell’uomo: Cos’è persona? Cosa sono i diritti? ~~– Comune di Gorgonzola Vai ai contenuti Vai al footer Regione Lombardia Accedi Comune di Gorgonzola Seguici su Youtube Youtube Cerca Cerca Apri burger menu Nascondi la navigazione Chiudi burger menu Comune di Gorgonzola Amministrazione Novità Servizi Vivere il Comune Tutti gli argomenti... Home /~~ Vivere il Comune / Eventi / Evento di formazione / I diritti dell'uomo: Cos'è persona? Cosa sono i diritti? I diritti dell'uomo: Cos'è persona? Cosa sono i diritti? 24 ottobre Primo incontro di un percorso per giovani attraverso esperienze e testimonianze ~~Condividi Facebook Twitter Linkedin Whatsapp Telegram Vedi azioni Stampa Ascolta Invia Vai al calendario eventi Indice della pagina Cos'è A chi è rivolto Luogo Date e orari Costi Allegati Contatti Ulteriori informazioni Cos'è~~ Con l’educatore ____ e don ____ ci porremo alcune domande fondamentali all’inizio di questo percorso sui diritti dell’uomo per un confronto e un dibattiro di sviluppi futuri. A chi è rivolto A tutti a ragazzi e ragazze tra i 17 e i 18 anni in su. Luogo Centro Intergenerazionale Via Italia, 84 ~~Gorgonzola MI, 20064 + – Leaflet~~ Date e orari 24 Ott Inizio ore 21:00 Costi Partecipazione gratuita Ingresso gratuito Allegati LOCANDINA I diritti dell'uomo Contatti …

Pulizia dei testi estratti

Approccio euristico:

Creazione di una «blacklist» di n-gram (singole parole, bigram e trigram nel caso sperimentale) che appaiono in più del 95% delle pagine di un sito

+

Pulizia tramite regex (espressioni regolari)

Prompt engineering

Promemoria: «Arte di guidare la conversazione con un agente conversazionale artificiale» (specificità, semplicità, esemplificazione, …)

Zero-Shot	Data Augmentation	Feature Extraction
`Dato il seguente documento: «{doc}». Quali categorie sono adatte a descriverlo tra le seguenti? «{tags}». Riporta solo le categorie, da un minimo di 1 a un massimo di 3, separate da una virgola.`	`Data la seguente pagina a tema «{label}», genera una nuova pagina con lo stesso tema. Pagina: «{doc}».`	`Trova quali categorie nella lista «{features}» sono presenti nel seguente documento: «{doc}». Ritorna solo le categorie, separate da una virgola.`

Note sulla classificazione supervisionata

Negli approcci Data Augmentation e Feature Extraction non facciamo altro che produrre nuovi testi fittizi e estrarre informazioni dai testi.

La classificazione è effettuata tramite un modello supervisionato, addestrato sull'80% dei dati (training set) e testato sul 20% (validation set / test set).

A questo fine è stato usato Born*, un modello supervisionato scelto per la sua velocità e capacità di gestire tutti i casi in esame.

Data Augmentation	Feature Extraction
TFIDF dei testi di training + Modello supervisionato	Feature binarie da linee guida AGID (es: presenza di nomi di persone, indirizzi, biografie, date, compensi, ...) + Modello supervisionato

* https://github.com/eguidotti/bornrule

Post-processing

Problema 4: Similarità/ambiguità tra tag diversi

Urbanizzazione

Similarità morfo-semantica

Urbanistica

Patrimonio culturale

Contenimento esplicito

Cultura e patrimonio culturale

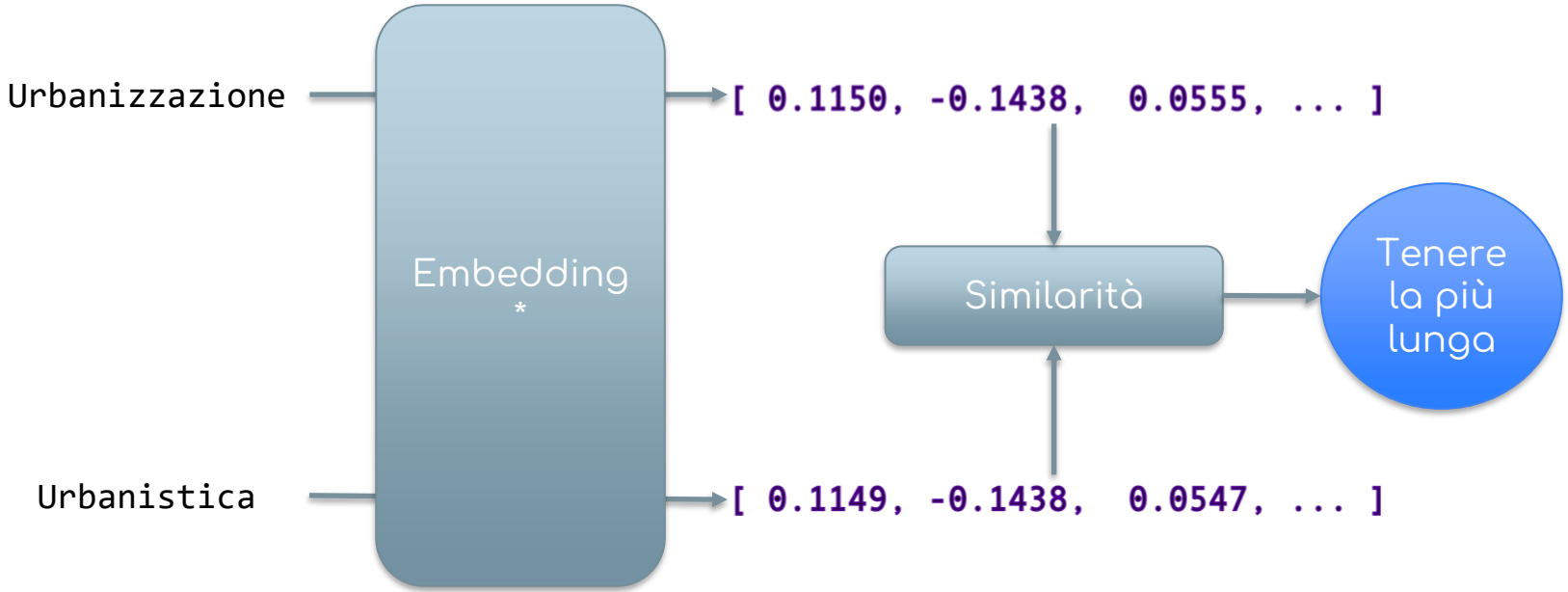
Concerto

Contenimento semantico

Manifestazione musicale

Post-processing

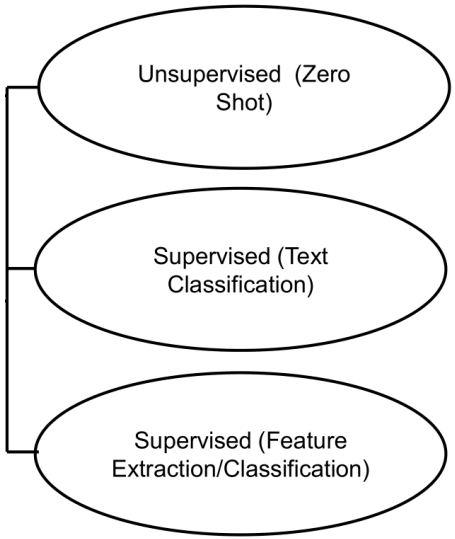
Soluzione proposta: Mappa semantica



* https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

Parte 2
Risorse necessarie

Risorse



LLM

Modello in cloud

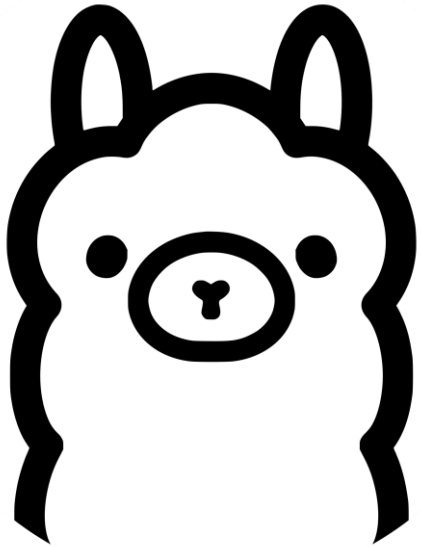
- Pay-per-use*
- Nessun requisito sul computer locale
- Dati trasmessi a terzi

Deployment locale

- Gratuito
- Requisiti tecnici dipendenti dai modelli utilizzati
- Dati in locale

* Difficile quantificazione precisa ma per gli esperimenti effettuati nell'ordine di una decina di euro utilizzando il modello meno oneroso.

Risorse



OLLAMA

Deployment locale

- Gratuito
- Requisiti tecnici dipendenti dai modelli utilizzati
- Dati in locale

Ollama è un software che permette l'esecuzione di LLM open-source in locale con poche semplici righe di codice, senza preoccuparsi dei dettagli tecnici.

Risorse

Esempi di requisiti di memoria RAM per deployment locali	GB
Gemma 2 (2b)	7
Phi-3.5 (3.8b)	9
Mistral (7b)	17
Mixtral (8x22b)	104
LLaMa 3.3 (70b)	172

Tool quali Ollama (https://ollama.com/), VLLM (https://docs.vllm.ai/en/latest/), llama.cpp (https://llama-cpp-python.readthedocs.io/en/latest/) e Gpt4All (https://www.nomic.ai/gpt4all) facilitano l'installazione locale degli LLM più diffusi, ottimizzando anche la loro impronta in memoria grazie a meccanismi quali Flash Attention, Paged Attention, Quantizzazione, ecc.

Parte 3
Dati raccolti

Comuni selezionati

- Condizioni:
  - tagging almeno parzialmente effettuato
  - consenso all'attività di scraping

- Dataset finale
  - Gorgonzola (MI)
  - Mandello del Lario (LC)
  - Pioltello (MI)
  - Tirano (SO)

https://comune.mandello.lc.it/it-it/novita/avvisi/2024/sociale/giovani-competenti-junior-per-comunita-16-20-anni-339867-1-889c1a2cf013adebf9d7e6fd7b6e223b

Argomenti (rilevati: 41)

Accesso all'informazione, Tempo libero, Turismo, Comunicazione istituzionale, Sport, Igiene pubblica, Assistenza agli invalidi, Assistenza sociale, Risposta alle emergenze, Associazioni, Gestione rifiuti, Piano di sviluppo, Sviluppo sostenibile, Urbanizzazione, Istruzione, Commercio ambulante, Spazio Verde, Formazione professionale, Imprese, Residenza, Concorsi, Protezione civile, Lavoro, Commemorazione - Ricorrenza, Mercatino dell'antiquariato, Incontro, Manifestazione musicale, Spettacolo teatrale, Festa, Manifestazione sportiva, Politica commerciale, Mercato, Imposte, Esposizione - Rassegna, Biblioteca, Proiezione cinematografica, C.I.M., Iniziativa per bambini e ragazzi, Partecipazione, Cultura e patrimonio culturale, Nascita

Categorie (rilevate: 12)

Allerte Meteo - Protezione Civile, Commercio e Mercati, Cultura, Eventi e Manifestazioni, Servizi sociali, Servizi al cittadino, Ambiente, territorio e urbanistica, Comunicazione, Edilizia, Avvisi, Contributi, Giovani, Bandi

Statistiche dataset finale

Numero medio di pagine estratte per Comune	89
Numero medio di pagine taggate con «Argomenti» per Comune	24
Numero medio di pagine taggate con «Categoria» per Comune	5
Lunghezza media testo pulito della pagina (in parole)	845
Numero medio di tag «Argomenti» per pagina	2
Tag «Argomenti» più frequenti	1. Tempo libero (30) 2. Accesso all'informazione (22) 3. Cultura e patrimonio culturale (17)
Tag «Categoria» più frequenti	1. Servizi sociali (5) 2. Avvisi (5) 3. Commercio e mercati (4)

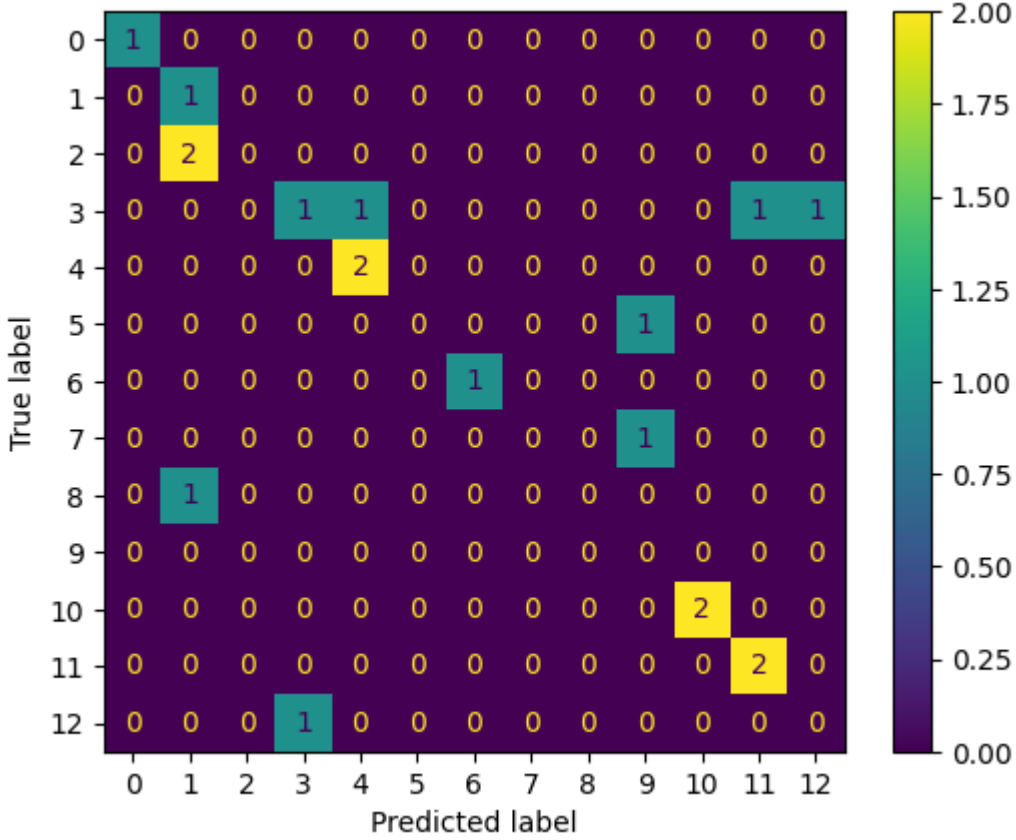
Parte 4
Risultati ottenuti

Risultati
Matrice di confusione

La matrice di confusione mette in evidenza gli errori.

La figura riguarda il modello zero-shot multiclass con Mistral come LLM.

0. Allerte Meteo – Protezione Civile
1. Commercio e Mercati
2. Cultura, Eventi e Manifestazioni
3. Servizi sociali
4. Servizi al cittadino
5. Ambiente, territorio e urbanistica
6. Comunicazione
7. Edilizia
8. Avvisi
9. Contributi
10. Giovani
11. Bandi
12. Nessuna delle precedenti



Risultati
Classificazione per Argomenti (multilabel)

Baseline randomica: 2.4%

	Precision	Recall	F1
Zero-Shot (GPT-4o-Mini)	47-48%	65-77%	54-56%
Zero-Shot (Mistral)	8-14%	57-64%	13-23%
Zero-Shot (Phi-3.5)	7-15%	32-34%	10-20%
Zero-Shot (Gemma)	25-27%	14-17%	18-19%
Data Augmentation (Mistral)	-	-	-
Feature Extraction (Mistral)	4%	47%	7%

Risultati
Classificazione per Categoria (multiclass)

Baseline randomica: 8.3%

	Precision	Recall	F1
Zero-Shot (GPT-4o-Mini)	79-82%	71-74%	72-78%
Zero-Shot (Mistral)	75-79%	69-72%	69-75%
Zero-Shot (Phi-3.5)	79-82%	66-67%	68-73%
Zero-Shot (Gemma)	-	-	-
Data Augmentation (Mistral)	52-53%	13-17%	21-25%
Feature Extraction (Mistral)	83-84%	20-21%	33-34%

Risultati
Classificazione per Argomenti (multilabel)

Tag argomenti in eccesso:

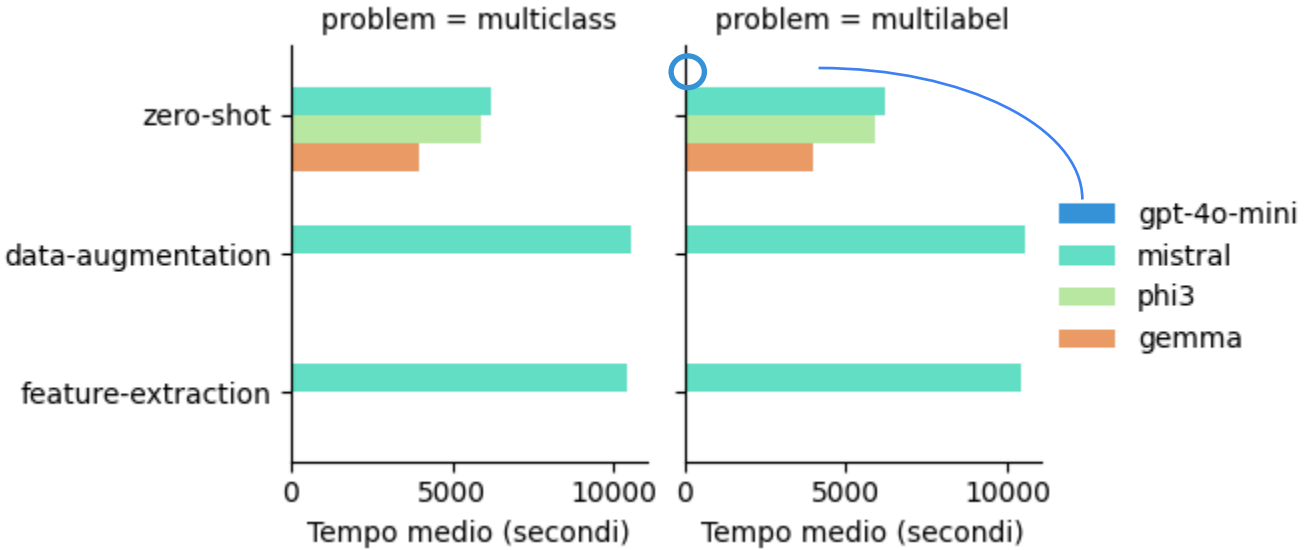
- Comunicazione istituzionale

Risultati
Classificazione per Argomenti (multilabel)

Tag argomenti in eccesso:

- Comunicazione istituzionale
- Urbanizzazione
- Eventi

Risultati
Tempi di calcolo (RAM 32 GB, senza GPU)



Conclusioni

1. Permane il problema della scarsità dei dati annotati (pagine web taggate), e sia l'approccio «Feature Extraction» sia l'approccio «Data Augmentation» non sembrano efficaci nell'affrontarlo

2. L'utilizzo di modelli open-source con deployment locali richiede un'architettura più consona per eguagliare i tempi di calcolo di modelli in cloud quali ChatGPT, ma riesce parzialmente a competere rispetto all'accuratezza delle previsioni

3. Estensione dell'attività: espansione del dataset di riferimento, affinamento delle fasi di pre-processing e post-processing, scelta del modello LLM e di un'infrastruttura

Grazie per l'attenzione!

Domande?

Contatti:
davide.riva1@unimi.it
marzio.decorato@unimi.it

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
MUSA
Missione 4 - Istruzione e ricerca