ATECO
CLASSIFICAZIONE
ATTIVITÀ ECONOMICHE

09 OTTOBRE 2025

Dati testuali e intelligenza artificiale:
l'uso di bilanci d'impresa e dati di indagine per la riclassificazione dei codici ATECO

Giulio Massacci
Donato Summa
Direzione Centrale per la Metodologia e il Disegno dei Processi Statistici

Istat | Istituto Nazionale di Statistica

Sommario

Contesto e obiettivo

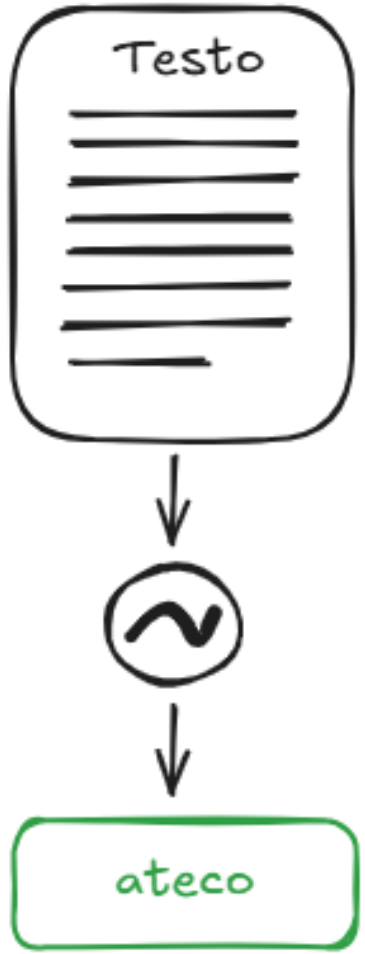
Contesto

La classificazione delle imprese **richiede** sempre più di integrare, **oltre agli indicatori economico-finanziari** tradizionali, **l'analisi di dati testuali** che ne descrivono caratteristiche e attività.

Obiettivo del progetto

Il lavoro ha avuto l'obiettivo di **valutare le potenzialità** di modelli di analisi testuale applicati a **fonti informative eterogenee**, al fine di supportare **processi più accurati** di classificazione.

- Supporto costruzione tabella operativa di riclassificazione ATECO 22-ATECO 25
- Gestione casi di non automatica codifica ATECO 25
- Acquisizione ed analisi di ulteriori fonti dati



Fonti dati testuali a disposizione 1/2

Bilanci

Oltre a informazioni numeriche, nei bilanci sono presenti le **note integrative** che forniscono una **descrizione dettagliata** e discorsiva della situazione economica dell'azienda.

Pro: **dichiarazione esplicita** delle attività principali dell'azienda, spesso in riferimento alla classificazione ATECO.

Contro: descrizione attività **non sempre presente** ed inserita come passaggio isolato all'interno di un testo esteso su **tematiche differenti**.

Indagine

Le informazioni testuali emergono soprattutto nei casi in cui l'indagine **non riesce a rappresentare** tutte le situazioni e/o per **preferenza del rispondente**.

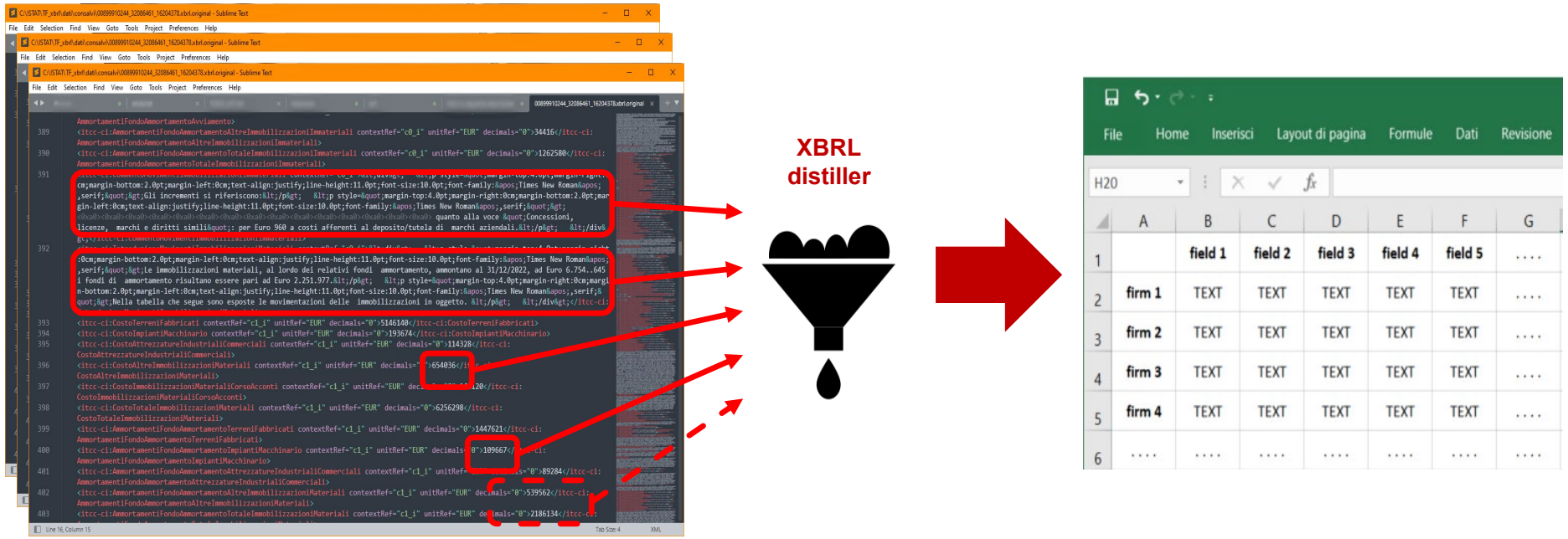
Pro: **individuazione dei testi** in cui è presente la dichiarazione delle attività data la natura schematica dell'indagine.

Contro: la **libertà di risposta** lascia a potenziali molteplici interpretazioni in combinazione con altre risposte dell'indagine.

Fonti dati testuali a disposizione 2/2

XBRL distiller è un software configurabile per **estrarre testi dei bilanci (formato XBRL) in forma tabellare** al fine di consentire successive analisi.

Workflow di estrazione automatica del testo dai bilanci delle imprese in formato XBRL



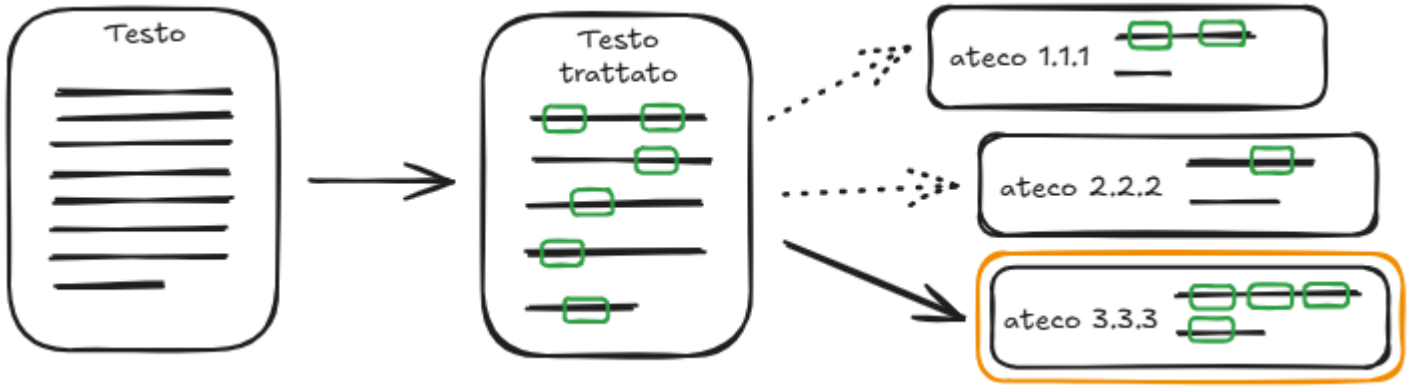
ATECO | Istat

Metodi tradizionali

Trattamento del testo con:

- **Rimozione** di parole ininfluenti
- **Trasformazione** delle parole nella loro forma radice
- **Inserimento** di sinonimi

Similarità tra testi mediante **conteggi di parole** in comune.

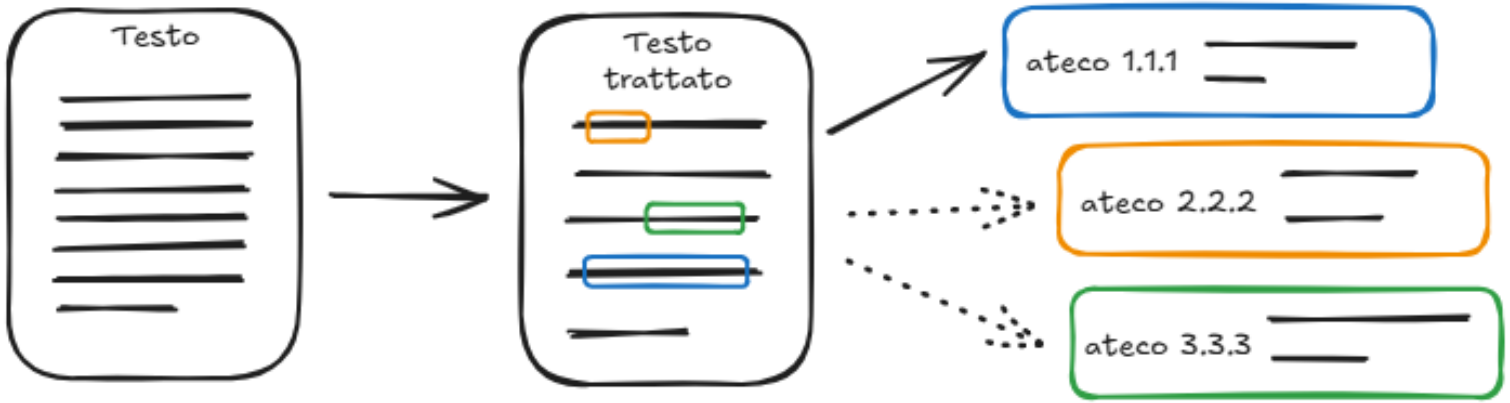


4%

Casi di corretta classificazione con ATECO 22

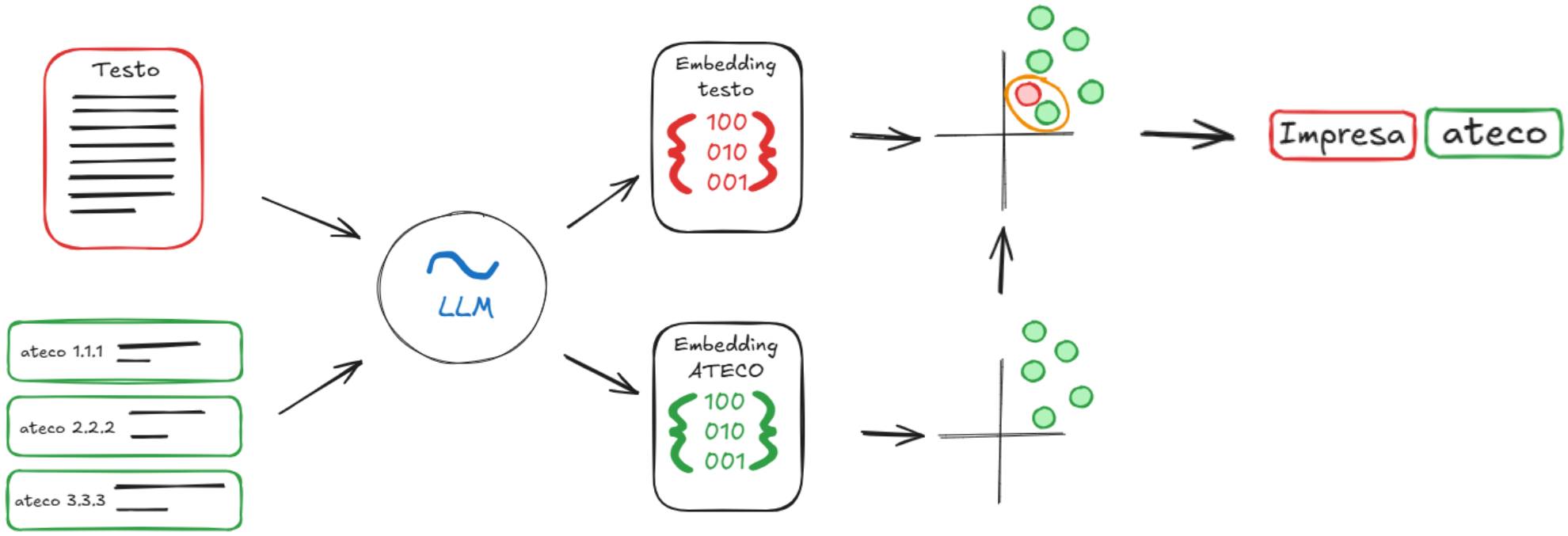
Metodi avanzati con l’uso Large Language Model (LLM) 1/2

Si rende **necessario un nuovo paradigma** di analisi semantica: dal semplice conteggio di parole condivise a una **comprensione più profonda dei contenuti**, resa possibile dai **modelli linguistici di ultima generazione** (LLM).



Metodi avanzati con l’uso Large Language Model (LLM) 2/2

Workflow di processamento dati testuali per associare l’impresa al codice ATECO.



Risultati (performance su ATECO 22)

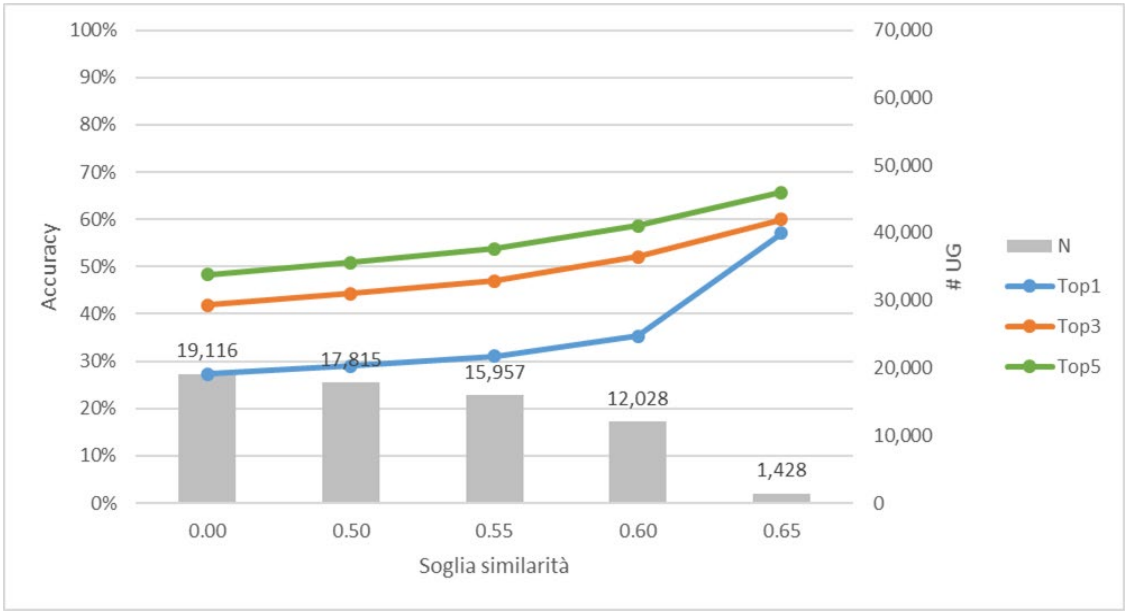
Il grafico mostra la **percentuale di casi di ATECO 22 classificati correttamente** (accuracy) al variare della soglia di similarità.

Viene riportata l'accuracy utilizzando la **prima scelta**, le **prime 3 scelte** e le **prime 5 scelte**.

Il modello ha un forte incremento di **accuracy** a partire da un **determinato valore soglia** (0.55) e suggerisce di **valutare l'assegnazione su più proposte** (da TOP3 in sù)

Campione ottenuto **stratificando per tipologia di imprese**.

PERFORMANCE DEL MODELLO NEL CLASSIFICARE LE IMPRESE IN ATECO 2022



Risultati (applicazione su ATECO 25)

I risultati sono analisi campionarie svolte da revisori. La stima di accuratezza è da prendere come indicazione.

Bilanci

59%

Casi corretti
sulla **prima scelta**

Con conoscenza ATECO 22 di partenza
Campione: 86 unità

Indagine

49%

Casi corretti
sulla **prima scelta**

Senza conoscenza ATECO 22 di partenza
Campione: 215 unità

Conclusioni e riflessioni

L’analisi di **dati testuali** sta assumendo un ruolo **sempre più centrale** nei processi di classificazione.

I **risultati attuali** evidenziano come questi modelli **rispondano ai requisiti della classificazione**, aprendo la strada a **una futura metodologia generalizzata** per l’uso delle stringhe nelle statistiche ufficiali.

Gli **sviluppi futuri** si focalizzeranno sulla verifica dell’**affidabilità dei dati testuali** e sul **perfezionamento dei modelli** che ne descrivono la struttura.

Non si tratta di un sostituto del lavoro, ma aiuta a **gestire scenari complessi** e a **supportare le decisioni finali**.

ATECO
CLASSIFICAZIONE
ATTIVITÀ ECONOMICHE

Istat | Istituto Nazionale di Statistica

09 OTTOBRE 2025

Dati testuali e intelligenza artificiale:
l'uso di bilanci d'impresa e dati di indagine
per la riclassificazione dei codici ATECO

Grazie

PER L'ATTENZIONE