PROGETTI IV CALL 2023/2024

Titolo progetto: **Applicazione di tecniche di web intelligence e modelli di intelligenza artificiale per la codifica automatica dei prodotti realizzati dalle imprese**

Descrizione:

L'Istat ha avviato un progetto innovativo orientato a migliorare la qualità e l'efficienza delle indagini statistiche sulle imprese, focalizzandosi sulla codifica automatica dei prodotti attraverso l'uso di tecniche di web scraping e modelli di intelligenza artificiale. L'obiettivo è sviluppare una soluzione scalabile e sostenibile che integri dati pubblici e tecnologie avanzate per supportare i processi di raccolta dati, ridurre i tempi di esecuzione e minimizzare i costi, garantendo al contempo il rispetto della privacy. Il progetto, sperimentato nell'ambito dell'Indagine Annuale sulla Produzione Industriale (ProdCom), mira a facilitare la classificazione automatica dei prodotti, migliorare la gestione delle risposte mancanti e preparare strumenti di supporto per gli utenti, con l'intento di estendere l'approccio a tutte le divisioni Ateco coinvolte. Potranno, infine, essere valutate le possibilità di applicazione dei risultati conseguiti ad altre rilevazioni statistiche che raccolgono dati sui prodotti.

Obiettivi:

- Incrementare l'efficienza dei processi di raccolta e la qualità dei risultati delle indagini statistiche sulle imprese che utilizzano i prodotti come unità di analisi, attraverso la codifica automatica dei prodotti riportati nei questionari e la corretta guida dei rispondenti nella compilazione.

- Estendere l'integrazione dei dati relativi ai prodotti a qualsiasi caso di mancata risposta totale e/o parziale.

- Contribuire alla predisposizione di un archivio statistico dei prodotti realizzati dalle imprese e progettare strumenti di codifica automatica da mettere a disposizione degli utenti per una corretta codifica (ad esempio, tramite applicazione web).

- Estendere l'attività sperimentale a tutte le divisioni Ateco rientranti nel campo di osservazione di ProdCom.

Metodologia:

L'impianto metodologico mira a sviluppare un sistema per la codifica automatica dei prodotti delle imprese, utilizzando tecniche di web intelligence e modelli di Intelligenza Artificiale (IA), con un focus primario sull'Indagine Annuale sulla Produzione Industriale (ProdCom) ma con potenzialità di estensione ad altre indagini sui prodotti.

La prima macro-fase è dedicata all'identificazione automatica dei siti web aziendali, partendo dal registro delle imprese, mediante tecniche di web scraping (utilizzando un software Java per scaricare contenuti testuali e identificare dati come indirizzi, email, P.IVA) e il successivo matching di tali dati con il Registro Statistico delle Imprese con analisi manuale per i casi di mancata individuazione.

La seconda macro-fase, dedicata all'estrazione e classificazione dei prodotti dai siti web aziendali, esplora due approcci principali, preceduti da un'analisi manuale mirata a identificare i diversi criteri di accesso alle pagine prodotto sui siti aziendali.

Il primo approccio sviluppato in linguaggio Python combina tecniche di web scraping con modelli di IA generativa. Questo prevede l'estrazione di nomi e descrizioni dei prodotti tramite scraper dinamici (come Scrapy/Playwright o librerie come scrapegraphai, supportati da LLM per strutturare l'output in JSON) e la successiva classificazione. Per la classificazione, sono state testate due strategie: (i) la prima basata su tecnologie open-source che utilizza LLM (Llama3.2:3B) per la riformulazione contestualizzata delle descrizioni dei prodotti successivamente trasformati in embedding tramite modelli open-source (Sentence Transformer) e confrontati tramite similarità semantica (distanza coseno) e LLM con le descrizioni trasformate in embedding delle voci della classificazione di riferimento al fine di individuare quella più attinente; (ii) la seconda basata su tecnologie proprietarie che utilizza LLM (OpenAI GPT-4o-mini) per la riformulazione contestualizzata delle descrizioni dei prodotti successivamente trasformati in embedding tramite modello proprietario OpenAI's text-embedding-3-large e salvati nel database vettoriale Qdrant utilizzato per contenere anche gli embedding delle descrizioni del sistema di classificazione di riferimento. L'abbinamento prodotto-codice viene effettuato tramite query semantiche sul database vettoriale.

Il secondo approccio è una strategia interamente open-source, sviluppata in linguaggio R. Inizia con una mappatura completa dei siti, estrazione e "chunking" del testo, trasformazione in vettori tramite modelli di embedding open-source (es. snowflake-arctic-embed2), e l'uso di un LLM open-source (Llama3.1:8B) per ottenere una lista di prodotti. Per la classificazione finale all'interno di questo approccio, si sono sperimentate due strategie: (i) l'impiego di un LLM specificamente addestrato (LabinLlama) sulla nomenclatura ProdCom; (ii) un'architettura ibrida che combina embeddings con modelli deep-learning (Variational Autoencoder-VAE integrato con tecniche di metric learning e focal loss). Quest'ultimo sistema VAE apprende uno spazio latente strutturato, ottimizzato per allineare gli embedding delle descrizioni dei prodotti con quelli dei codici ProdCom, gestendo anche lo squilibrio tra le classi, per poi classificare tramite similarità coseno.

Entrambi gli approcci mirano a migliorare l'efficienza delle indagini, ridurre i costi e l'onere statistico sulle imprese.

Risultati ottenuti:

L'integrazione tra continuità e innovazione è fondamentale nella produzione statistica: la continuità garantisce la comparabilità nel tempo, mentre l'innovazione permette di arricchire l'offerta informativa con nuovi prodotti. L'uso combinato di

tecniche di web intelligence e AI rappresenta un'importante opportunità per migliorare la qualità e la profondità delle informazioni statistiche a supporto delle politiche pubbliche.

L'applicazione sperimentale di due approcci basati su web scraping e modelli di linguaggio (LLM) per la classificazione automatica dei prodotti secondo il sistema ProdCom ha mostrato risultati preliminari promettenti.

In particolare, relativamente alla fase finale di classificazione si evidenzia che, con il primo approccio è stato possibile raggiungere un'accuratezza del 45% delle predizioni dei codici prodotto a otto cifre (verificate manualmente); per il 25% circa dei prodotti classificati in modo errato risultano comunque correttamente attribuite le prime sei cifre su un totale di otto. Con il secondo approccio, basato su un modello di deep learning addestrato e validato con dati ufficiali provenienti dalla tabella di conversione NC8- ProdCom, il grado di accuratezza, a livello di prodotto a otto digit, è stato del 51% e a livello di almeno sei digit del 22,95%, considerando i prodotti non classificati correttamente. Le metriche macro weigthed precision, recall e F1-score sono state rispettivamente pari al 71,27%, 51% e 69,84%. Entrambe le soluzioni – una basata su modelli proprietari e l'altra completamente open source – richiedono tuttavia ulteriori affinamenti mirati al consolidamento dei risultati.

In generale, le tecniche oggetto di sperimentazione offrono prospettive concrete per aumentare l'efficienza delle rilevazioni ufficiali, ridurre i costi e l'onere statistico e migliorare la qualità degli output. Tuttavia, l'implementazione è complessa e richiede gradualità, con una fase di transizione impegnativa nel breve periodo ma con potenziali benefici significativi nel medio termine.

Membri del Team:	*Amarone Massimiliano, Bianchi Gianpiero, Briatico Luigi, D'Amore Gabriele, Gambuti Teresa, Morrone Mirella, Moscufo Maria, Papa Pasquale, Pianura Paola, Scalfati Francesco, Summa Donato.*