Analysing Business Data with Statistics, Data Science, and AI brings together a selection of original contributions that explore how statistical methodology, data science, and artificial intelligence can jointly enhance the analysis and interpretation of business and socio-economic data.

The volume stems from the scientific discussions developed during the international conference "Measuring and Interpreting World Changes with Statistics, Data Science and AI", held in Rome from 18 to 20 September 2024. The conference was jointly organised by the Association for Applied Statistics (ASA), the Department of Statistical Sciences of Sapienza University of Rome, and the Italian National Institute of Statistics (ISTAT), with the participation of several academic and institutional partners. The event provided a multidisciplinary forum for examining the contribution of statistics, data science, and artificial intelligence to the understanding of contemporary economic and social transformations.

The Special Issue addresses both theoretical and applied perspectives, focusing on the growing availability of complex, high-dimensional, and heterogeneous data generated by firms, institutions, and digital platforms. Particular attention is devoted to the integration of traditional statistical sources with non-traditional data, including administrative records, digital traces, and textual information, as well as to the development of transparent and explainable AI models capable of supporting evidence-based decision making.

The contributions collected in this volume investigate advanced methodological frameworks, empirical applications, and interdisciplinary approaches aimed at improving data-driven strategies in areas such as business analysis, labour markets, innovation processes, and territorial dynamics. By combining classical statistical reasoning with modern machine-learning techniques, the volume highlights both the opportunities and the challenges associated with the adoption of AI-based tools, including issues related to interpretability, data quality, and responsible use of algorithms.

Intended for researchers, practitioners, and policymakers, this Special Issue provides a rigorous and coherent overview of current developments at the intersection of statistics, data science, and artificial intelligence, contributing to the ongoing scientific debate on how quantitative methods can effectively support decision-making processes in complex socio-economic contexts.

This volume is published within the TESI & TEMI editorial series, jointly promoted by Universitas Mercatorum and the Centro Studi delle Camere di Commercio G. Tagliacarne, and reflects their shared commitment—developed in cooperation with the Association for Applied Statistics—to fostering high-quality research and scientific dialogue in applied statistics and business analytics.

ISBN 978-88-9326-281-1

Special issue 1 **Analysing Business Data with Statistics, Data Science, and AI** TEMI

TEMI territori economie mercati istituzioni

Dipartimento di Scienze Statistiche Sapienza Università di Roma

ASA Associazione per la Statistica Applicata

Istat Istituto Nazionale di Statistica

Special issue 1

Analysing Business Data with Statistics, Data Science, and AI

Editors
Fabio Crescenzi, Luigi Fabbris, Andrea Mazzitelli, Alessandra Righi, Alessandro Rinaldi, Maurizio Vichi

Articles

1986 2026 40 ANNI

CENTRO STUDI DELLE CAMERE DI COMMERCIO GUGLIELMO TAGLIACARNE

Universitas Mercatorum
Università telematica delle Camere di Commercio Italiane

This Special Issue inaugurates a broader editorial initiative aimed at collecting and disseminating high-quality scientific contributions at the intersection of statistics, data science, and artificial intelligence. Conceived as part of an ongoing Special Issue programme, the volume reflects a structured editorial experience designed to evolve over time through further thematic issues.

The philosophy underlying the Special Issues is to promote continuity in scientific debate, methodological innovation, and interdisciplinary dialogue, while ensuring rigorous peer review and editorial coherence. By bringing together contributions developed within shared scientific contexts and extended through subsequent editorial projects, the series aims to support cumulative knowledge building and to respond to emerging analytical challenges in business, economics, and social research.

TEMI Territori Economie Mercati Istituzioni

Dipartimento di Scienze Statistiche
Sapienza Università di Roma

ASA Associazione per la Statistica Applicata

Istat Istituto Nazionale di Statistica

Special issue 1

Analysing Business Data with Statistics, Data Science, and AI

1986 2026 ANNI

Centro Studi delle Camere di Commercio Guglielmo Tagliacarne

Università Mercatorum
Università telematica delle Camere di Commercio Italiane

The joint digital editorial series promoted by Universitas Mercatorum and the Centro Studi delle Camere di Commercio G. Tagliacarne include:
TESI (Territory, Economy, Society, Institutions). Instant Paper: blog-based publications subject to a preliminary assessment of scientific coherence;
TESI (Territory, Economy, Society, Institutions). Paper: aperiodic publications without ISBN, reviewed through a single-blind peer review process;
TESI (Territory, Economy, Society, Institutions). Discussion Paper: aperiodic publications with ISBN assigned by Universitas Mercatorum, subject to double-blind peer review;
TEMI (Territory, Economy, Markets, Institutions): a series collecting theoretical and analytical contributions selected through thematic calls for papers addressing topics relevant to the scientific communities of Universitas Mercatorum and the Centro Studi delle Camere di Commercio G. Tagliacarne.

Distribution platforms:
https://www.tagliacarne.it/tesi_temi-30
https://www.unimercatorum.it/ricerca/tesi-e-temi

Aperiodic publication. Copyright © 2022 Proprietors and publishers:
Universitas Mercatorum
Piazza Mattei 10, 00186 Rome
Centro Studi delle Camere di Commercio G. Tagliacarne
Piazza Sallustio 9, 00187 Rome

Editor: Giapeto Editore srl con socio unico - Napoli
First edition: February 2026
ISBN: 978-88-9326-281-1

INDICE

EDITORIAL

Fabio Crescenzi[1], *Luigi Fabbris*[2], *Andrea Mazzitelli*[3], *Alessandra Righi*[4], *Alessandro Rinaldi*[5], *Maurizio Vichi*[6]

1. Introduction

The analysis of business data is undergoing a profound transformation driven by the increasing availability of large, heterogeneous and rapidly evolving data sources. Businesses and institutions now operate in information-rich environments where administrative data, transactional records, digital platforms, social media and textual repositories coexist and interact.

These developments require analytical frameworks capable of integrating structured and unstructured data, handling scale and complexity, and producing timely and actionable insights. At the same time, the growing reliance on automated and AI-based methods raises critical questions concerning interpretability, transparency and governance, especially when analytical results directly support business strategies or public decisions.

In this context, artificial intelligence should be interpreted as an extension of statistical reasoning rather than a substitute for it. Statistical principles remain essential to ensure that AI-driven analyses of business data are methodologically sound, reproducible and meaningful. This special issue brings together seven contributions that jointly explore how statistics, data science and AI can be combined to address concrete analytical challenges in business and institutional settings.

This special issue is part of a broader editorial initiative promoted by the Association for Applied Statistics (ASA) and Unioncamere – Centro Studi delle Camere di Commercio Guglielmo Tagliacarne through the *Special issues series* of the journal Tesi & Temi. The initiative originates from the international conference "Measuring and Interpreting World Changes with Statistics, Data Science and AI", held in Rome from 18 to 20 September 2024, jointly organised by ASA, the Department of Statistical Sciences of Sapienza University of Rome and the Italian National Institute of Statistics (ISTAT), together with other prestigious partners. The special issues aim to collect selected contributions presented at the conference, while also welcoming additional papers that

[1] Former Istat - Istituto Nazionale di Statistica, Rome, Italy - e-mail: fabio7826@gmail.com
[2] University of Padua, Padua, Italy - e-mail: fabbris@stat.unipd.it
[3] Universitas Mercatorum, Rome, Italy - e-mail: a.mazzitelli@unimercatorum.it
[4] Istat - Istituto Nazionale di Statistica, Rome, Italy - e-mail: righi@istat.it
[5] Centro Studi Tagliacarne, Rome, Italy - e-mail: alessandro.rinaldi@tagliacarne.it
[6] Sapienza University of Rome, Rome, Italy - e-mail: maurizio.vichi@uniroma1.it

explore the use of statistics, combined with technology and artificial intelligence, to support decision-making processes in public and private organisations. This special issue represents the first volume of the *Special issues series*, conceived as a multi-volume editorial project. Further special issues will follow, continuing to investigate the contribution of statistics, data science and artificial intelligence to decision support across different application domains.

2. Explainability and language-based models for business data

The increasing use of textual data in business analytics – ranging from customer feedback and online reviews to corporate reports and institutional documents – has made explainable language-based models a key requirement. In the paper "Textual Classification Explained by Counterfactual Analysis in LLMs", Mauro Sodani and Valerio De Camillis address this challenge by proposing an interpretable framework for text classification based on large language models.

Their approach treats text generation as a probabilistic classification task and exploits token-level probability distributions, perplexity measures and counterfactual analysis. By systematically removing individual words and observing their impact on model confidence, the method identifies which textual elements drive classification outcomes.

This contribution is particularly relevant for business and institutional contexts in which textual classifications inform strategic decisions, regulatory monitoring or quality assessments, and explainability is essential to ensure trust and accountability.

3. Conceptual foundations of artificial intelligence in business analytics

In "Decostruire l'IA: Tra Paure Pubbliche e Fondamenti Scientifici", Tonio Di Battista offers a conceptual reflection that is highly relevant for the interpretation of AI-based business analytics. The paper challenges widespread narratives that attribute autonomous intelligence to AI systems, arguing instead that contemporary AI is fundamentally grounded in statistical inference and computational optimization.

By reconnecting machine learning and deep learning methods to their statistical roots, the contribution provides a critical framework for understanding both the potential and the limitations of AI in business decision-support systems. This perspective helps prevent over-reliance on automated outputs and reinforces the role of human judgement and statistical reasoning in interpreting analytical results.

4. Ontologies and semantic structuring of online business-related contents

The growing importance of digital platforms in shaping markets, consumer behaviour and public debates has increased the relevance of methods for structuring and interpreting online textual contents. In "Learning Ontologies of Online Abusive Con-

tents: Seeded LDA and Graph-Based Semantic Structuring of Offensive Anti-migrant Narratives in Italian", Alex Cucco, Lara Fontanella, Annalina Sarra and Sara Fontanella propose a data-driven approach to ontology learning based on seeded topic models and semantic networks.

Although applied to online debates on migration, the methodology has broader implications for business analytics, where understanding narratives, sentiments and discursive patterns is essential for reputation management, market analysis and risk assessment. Ontology-based representations act as an intermediate layer between raw textual data and interpretable analytical outputs, enhancing explainability and contextual understanding.

5. Automatic synthesis of scientific and technical knowledge for business decisions

Businesses and institutions increasingly rely on scientific and technical knowledge to inform strategic decisions, innovation processes and policy design. In "Synthesizing Knowledge: An Integrated Approach for Extracting Relevant Content from Scientific Literature", Massimo Aria, Corrado Cuccurullo, Luca D'Aniello, Michelangelo Misuraca and Maria Spano address the challenge of information overload through an extractive summarisation approach tailored to structured scientific texts.

By integrating document structure, author-provided keywords and statistically grounded relevance measures, the proposed method enables transparent and reproducible synthesis of large document collections. Such tools are particularly valuable for competitive intelligence, technology scouting and evidence-based business strategy.

6. Digital contents, diffusion and virality in business and social environments

Understanding how information spreads in digital environments is crucial for marketing, reputation management and policy communication. In "Insights from Italian Tweets: Distributions, Content, Sentiment, Multimedia, and Network Metrics", Domenica Fiordistella Iezzi, Roberto Monte and Domenico Pasquini analyse large-scale Twitter data to investigate the statistical properties of viral messages.

The paper combines distributional analysis, sentiment indicators, multimedia features and network metrics to characterise highly retweeted contents, providing insights into the mechanisms that drive online popularity and engagement. Such analyses support data-driven strategies in digital business and communication contexts.

Complementarily, the paper "Data Science and AI: A Technology Proposal to Improve Statistical Innovation Processes" by Francesco Altarocca and co-authors proposes a low-code data science framework aimed at integrating administrative and business registers. This contribution highlights how technological infrastructures can facilitate the adoption of advanced analytics in business and institutional settings.

7. Responsible AI and governance of business-related statistics

In "Responsible AI Adoption: How It's Changing Official Statistics", Gerarda Grippo and Alessandra Righi examine the governance implications of adopting AI and machine learning in official statistics. While focused on institutional contexts, the issues discussed – data quality, algorithmic bias, transparency and accountability – are directly relevant for business analytics.

The paper emphasises that responsible AI adoption requires robust governance frameworks, adequate digital infrastructures and multidisciplinary skills. These conditions are essential to ensure that AI-driven analyses of data remain reliable, fair and aligned with organisational and societal objectives.

8. Concluding remarks and future directions

The seven contributions included in this special issue collectively show that analysing business data with statistics, data science and artificial intelligence requires more than computational power. It requires conceptual clarity, methodological rigour and institutional awareness.

Across different domains and applications, a common message emerges: statistics provides the unifying framework that allows heterogeneous data sources, complex models and automated procedures to be integrated into coherent and interpretable analytical processes.

Looking ahead, future research should further explore hybrid approaches that combine statistical modelling, machine learning and explainable AI, with a strong focus on business relevance and decision support. Strengthening the dialogue between methodological innovation and practical analytical needs will be essential to fully exploit the potential of AI while preserving trust, transparency and responsibility in business and institutional analytics.

CLASSIFICAZIONE TESTUALE SPIEGABILE TRAMITE ANALISI CONTROFATTUALE NEGLI LLM

TEXTUAL CLASSIFICATION EXPLAINED BY COUNTERFACTUAL ANALYSIS IN LLMS

Valerio De Camillis[1]*, Mauro Sodani*[2]

Sommario

La spiegabilità è essenziale per migliorare la trasparenza e la comprensione dei modelli di intelligenza artificiale in particolare nelle statistiche ufficiali. Nel caso di studio vengono analizzati i feedback degli utenti per valutare la classificazione automatica effettuata dall'IA.

Proponiamo una nuova metodologia per la classificazione e l'interpretazione di commenti testuali tramite modelli linguistici di grandi dimensioni, che integra l'analisi delle distribuzioni di probabilità sui token con tecniche controfattuali e misure di perplessità. La nostra soluzione consiste nell'utilizzare un prompt strutturato per classificare ciascun commento in uno spazio vettoriale a cinque dimensioni, corrispondente alle categorie semantiche di interesse, estraendo le probabilità associate ai token di categoria generati dal modello. Per valutare la rilevanza delle singole parole nel processo decisionale della LLM, rimuoviamo iterativamente ciascuna parola dal commento e misuriamo la variazione della perplessità nella classificazione. Questo approccio consente di identificare i termini più influenti e di analizzare la robustezza delle predizioni del modello. Confrontiamo la nostra tecnica con metodi di interpretazione consolidati come LIME e SHAP, evidenziando come il nostro metodo costituisca una categoria specifica progettata esplicitamente per l'analisi testuale. La metodologia proposta si ispira e si integra con recenti contributi sulla classificazione token-level, sulla stima dell'incertezza e sulle spiegazioni controfattuali.

Abstract

Explainability is essential to improve transparency and understanding of artificial intelligence models especially in official statistics. In the case study, user feedback is analysed to evaluate the automatic classification performed by AI.

We propose a new methodology for the classification and interpretation of textual comments using large language models (LLM), which integrates the analysis of

[1] Istat - Istituto Nazionale di Statistica, Roma, Italia - e-mail: decamillis@istat.it

[2] Istat - Istituto Nazionale di Statistica, Roma, Italia - e-mail: sodani@istat.it

probability probability distributions on tokens with counterfactual techniques and perplexity measures. Our solution consists of using a structured prompt to classify each comment in a five-dimensional vector space corresponding to the semantic categories of interest, extracting the probabilities (logprobs) associated with the category tokens generated by the model. To assess the relevance of individual words in the LLM decision-making process, we adopt a counterfactual procedure: we iteratively remove each word from the comment and measure the change in perplexity in the classification. This approach allows us to identify the most influential terms and to analyse the robustness of the model's predictions.

We compare our technique with established interpretation methods such as LIME and SHAP, highlighting how our method constitutes a specific category designed explicitly for textual analysis. The proposed methodology is inspired and complemented by recent contributions on token-level classification, uncertainty estimation in LLMs and counterfactual explanations.

Parole chiave: XAI (Intelligenza Artificiale Spiegabile), LIME (Spiegazioni Locali Interpretabili Modello-Agnostiche), SHAP (Spiegazioni Additive di Shapley), Controfattuali, Perplessità.
Keywords: *XAI (Explainable Artificial Intelligence), LIME (Local Interpretable Model-agnostic), SHAP (Shapley Additive explanations), Counterfactuals, Perplexity.*

1. Introduzione

In questo lavoro presentiamo una metodologia innovativa per la classificazione automatica di commenti testuali tramite modelli linguistici di grandi dimensioni (LLM), che integra l'analisi della distribuzione delle probabilità sui token con una tecnica controfattuale guidata dalla perplessità, per identificare le parole più rilevanti nei feedback. La tecnica si basa sull'idea di trattare la generazione del primo token dopo un prompt strutturato ({feedback} : {classificazione}) come un problema di classificazione a scelta multipla, dove ciascuna categoria corrisponde a un asse di uno spazio vettoriale a N dimensioni (nel caso di studio ISTAT N = 5 : "anomalie", "complimenti", "domande", "negativi", "osservazioni"). Per ogni commento, si estrae la probabilità (logprobs) associata al primo token di ciascuna categoria, ottenendo così una rappresentazione probabilistica continua della posizione del commento nello spazio semantico delle categorie.

Questa impostazione si ispira alle recenti ricerche che propongono di trattare la generazione di token nei LLM come un problema di classificazione, sfruttando pienamente le informazioni probabilistiche a livello di token per migliorare accuratezza e robustezza delle predizioni (Yu, *et al*., 2024). In particolare, la tecnica GaC (Generation as Classification) dimostra che l'uso diretto dei vettori di probabilità di classificazione

permette di prevenire errori a cascata nella generazione testuale e di ottenere performance superiori rispetto ai metodi di ensemble tradizionali basati sull'output testuale completo.

Parallelamente, il nostro approccio si inserisce nel filone degli studi sull'incertezza nei LLM, che evidenziano i limiti delle strategie basate sulla sola probabilità normalizzata per stimare la confidenza del modello. Come discusso da Thuy e Benoit (2024), la normalizzazione delle probabilità comporta una perdita di informazione sulla forza dell'evidenza accumulata dal modello durante l'addestramento, rendendo la probabilità un indicatore solo relativo e non assoluto dell'affidabilità della risposta.

A questa metodologia abbiamo aggiunto un passaggio controfattuale che consiste nel rimuovere, per ogni feedback, una parola alla volta e chiedere alla LLM una nuova classificazione. La variazione della distribuzione delle probabilità di output viene utilizzata come indicatore quantitativo della rilevanza della parola rimossa. Questo approccio si ispira al lavoro di Liu *et al*. sulla generazione di spiegazioni controfattuali model-agnostic, che mostrano come modifiche minime agli input possano essere utilizzate per spiegare le decisioni del modello e identificare le caratteristiche più influenti. Inoltre, recenti studi come quello di Cui *et al*. (2025) propongono di utilizzare la perplessità (definita nel paragrafo seguente) come misura dell'importanza di singoli elementi dell'input, ad esempio rimuovendo passaggi di ragionamento e osservando la variazione nella confidenza del modello (Stepwise Perplexity-Guided Refinement for Efficient Chain-of-Thought Reasoning in Large Language Models).

In letteratura, la generazione di spiegazioni controfattuali è stata approfonditamente studiata anche in termini di diversità, fattibilità e azionabilità delle modifiche proposte. Mazzine e Martens (2021) sottolineano che non esiste un singolo algoritmo migliore per generare spiegazioni controfattuali, poiché le prestazioni dipendono fortemente dalle proprietà del dataset, del modello e delle specificità del punto fattuale. Laugel *et al*. (2023) propongono una revisione delle numerose definizioni di diversità nelle spiegazioni controfattuali, categorizzandole in base a dimensioni come esplicito vs implicito e l'universo in cui sono definite. Stepin *et al*. (2021) forniscono una panoramica completa dei metodi contrastivi e controfattuali per l'IA spiegabile, evidenziando come l'incorporazione della contrastività migliori la qualità delle spiegazioni. Nel nostro lavoro, la rimozione iterativa di singole parole dal testo del commento genera una varietà di scenari controfattuali che possono essere analizzati sia in termini di impatto locale (sulla classificazione e sulla perplessità) sia in termini di diversità delle possibili spiegazioni ottenibili.

In un confronto con le metodologie più consolidate di interpretazione dei modelli, come LIME e SHAP, il nostro approccio presenta analogie e differenze significative. LIME (Local Interpretable Model-agnostic Explanations) costruisce un modello surro-

gato locale e lineare attorno alla predizione di interesse, perturbando le features di input e osservando le variazioni dell'output, mentre SHAP (SHapley Additive exPlanations) utilizza i valori Shapley della teoria dei giochi per attribuire un contributo a ciascuna feature nella predizione. Entrambe le tecniche mirano a spiegare le decisioni del modello identificando le features più rilevanti, ma operano tipicamente su dati tabellari e sono state successivamente adattate per l'applicazione a dati testuali. Nel nostro caso, invece, proponiamo una metodologia progettata specificamente per l'analisi testuale, che lavora direttamente sul testo non strutturato e sfrutta la capacità intrinseca dei LLM di fornire una misura di confidenza (perplessità) su ogni possibile output, estendendo così la logica di perturbazione e attribuzione delle spiegazioni ai modelli linguistici moderni in modo nativo. Questo approccio permette di ottenere spiegazioni locali e interpretabili anche in contesti complessi e ad alta dimensionalità, come è intrinsecamente lo spazio latente di una rete neurale profonda, superando alcune limitazioni di scalabilità e stabilità che possono emergere nell'applicazione di LIME e SHAP ai LLM.

L'implementazione utilizza un approccio many-shot in-context learning, ispirato ai recenti sviluppi della ricerca sui modelli linguistici (Argawal, *et al*., 2024), che dimostra come l'utilizzo di centinaia o migliaia di esempi nel contesto possa migliorare significativamente le prestazioni rispetto ai tradizionali approcci few-shot. Per il nostro caso d'uso di classificazione testuale, la ricerca suggerisce che l'utilizzo di 125-500 esempi per categoria potrebbe rappresentare un numero ottimale di "shots" per massimizzare le prestazioni di classificazione, come evidenziato negli esperimenti su task di sentiment analysis e classificazione di documenti.

La crescente attenzione alla spiegabilità e interpretabilità dei modelli si riflette anche nello sviluppo di metodologie per la generazione di spiegazioni controfattuali, che consentono di identificare quali modifiche agli input porterebbero a una diversa predizione, favorendo così una comprensione più profonda delle logiche decisionali dei modelli (Liu, *et. al*., 2025; Mothilal *et. al*., 2020; Chatterjee, *et al*., 2025). Nel nostro sistema, l'analisi controfattuale guidata dalla perplessità consente di mappare ogni commento in uno spazio continuo e di identificare le parole più rilevanti per la decisione del modello, migliorando ulteriormente la trasparenza e la facilità di analisi delle decisioni automatiche.

2. Perplessità

La perplessità è una metrica comune per gli LLM. Intuitivamente misura quanto un modello linguistico è incerto nel prevedere la parola successiva in una frase; un valore basso indica che il modello assegna probabilità elevate alle sequenze reali (è "poco perplesso").

Misura in modo quantitativo il grado di incertezza con cui un modello assegna probabilità a una sequenza di token data. In termini operativi, la perplessità è calcolata sulla

sequenza di output prevista dal modello in risposta a un prompt specifico, ovvero su una sequenza di token y = (y1,y2,…,yN) generata o valutata condizionatamente al prompt iniziale x.

Il calcolo si basa sulla probabilità congiunta assegnata dal modello a ogni token nella sequenza, considerando il contesto formato dal prompt e dai token precedenti:

$$\mathrm{PPL}(x, \{w_k\}_{k=1}^{N}) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i \mid x, w_1, \ldots, w_{i-1})\right)$$

dove:

- N è la lunghezza totale della sequenza di token
- $\{w_k\}_{k=1}^{N}$) indica la sequenza di token con lunghezza totale N che sono condizionati da x
- La probabilità p(wi | x, w1, w2, . . . , wi−1) è la probabilità condizionata assegnata dal modello al token wi-esimo dato il prompt x e i token precedenti
- l'esponenziale trasforma l'entropia media (in nats) in una scala più interpretabile, dove il valore minimo teorico è 1 (incertezza nulla, modello perfettamente sicuro).

Questa metrica è applicata esclusivamente alla sequenza target (o generata) che segue il prompt, non all'intero corpus né al prompt stesso. Tale scelta riflette l'obiettivo principale della perplessità: valutare quanto bene il modello "comprende" e "predice coerentemente" il proseguimento del testo in un contesto dato.

La rilevanza della perplessità nel contesto degli LLM è duplice:

1. Diagnostica: un valore basso indica che il modello assegna probabilità elevate ai token osservati, suggerendo una buona capacità di generalizzazione e coerenza linguistica.
2. Comparativa: permette di confrontare modelli diversi (o diverse configurazioni dello stesso modello) su uno stesso insieme di dati di test, offrendo una misura standardizzata dell'efficacia predittiva.

Nel nostro caso la utilizziamo per il suo potere comparativo fra prompt diversi, utilizzando ogni prompt come classificazione di un singolo feedback.

Sebbene la perplessità non catturi aspetti qualitativi complessi come la coerenza argomentativa o la correttezza fattuale, rimane uno strumento essenziale nella fase di sviluppo e ottimizzazione degli LLM, grazie alla sua natura matematica rigorosa e alla sua stretta connessione con la teoria dell'informazione. Fu introdotta originariamente

nel 1977 da Jelinek, Mercer, Bahl e Baker (1977) nel contesto del riconoscimento vocale, ed è oggi universalmente adottata come benchmark primario per la valutazione della qualità predittiva dei modelli linguistici.

3. Dati e Preprocessing

Il dataset impiegato è un ibrido di dati reali e sintetici. Partendo da un insieme di 500 commenti raccolti tramite piattaforme ufficiali di raccolta feedback degli utenti presso ISTAT. I commenti sono stati suddivisi manualmente in cinque categorie semantiche rilevanti: anomalie, complimenti, domande, negativi e osservazioni. La distribuzione approssimativa per categoria è la seguente: anomalie (20%), complimenti (25%), domande (15%), negativi (20%), osservazioni (20%). Sulla base di questi commenti un LLM è stato utilizzato per produrne altri di stile e qualità simile, fino ad arrivare ad un dataset totale di 2.500 commenti.

Per l'addestramento e la valutazione è stato utilizzato un criterio di suddivisione stratificata in training (70%) e test (30%) per garantire la rappresentatività di tutte le classi. Sono state adottate procedure di pulizia base, come la rimozione di duplicati, testi non pertinenti e caratteri non alfanumerici.

Tabella 1. Dataset

Origine	Commenti reali su siti web + commenti sintetici
Numerosità complessiva	500 reali + 2000 sintetici
Distribuzione	anomalie (20%), complimenti (25%), domande (15%), negativi (20%), osservazioni (20%)
Training/Test	70/30

Fonte: Elaborazioni a cura degli Autori

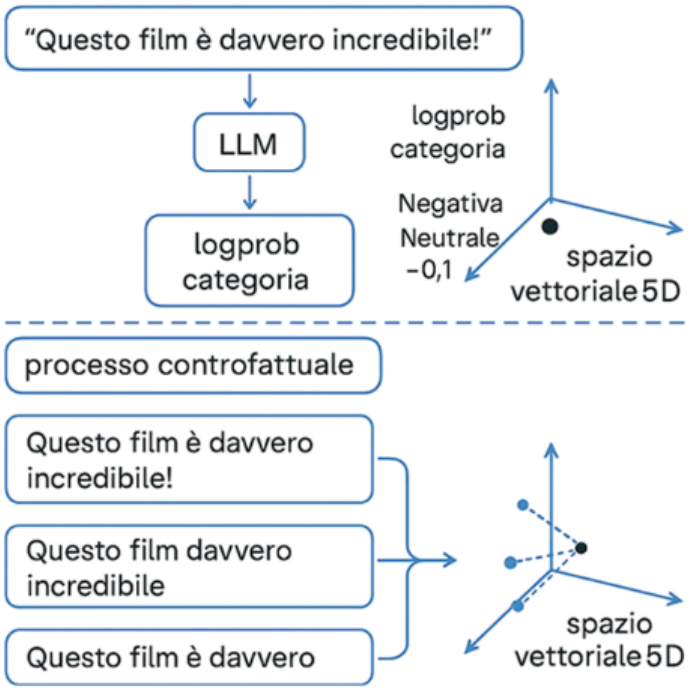
Questa divisione del dataset in training/test non è comunque necessaria per la tecnica di spiegabilità che presentiamo. Il nostro risultato è generalizzabile a qualunque modello pre-addestrato, quali LLM commerciali (OpenAi, Anthropic, etc.) o open weight (LLama, Deepseek, etc.). È stato necessario invece nella regressione logistica che rappresenta la baseline di confronto, come spiegato nel paragrafo 4.1.

4. Implementazione

L'implementazione della metodologia presentata nel paper è realizzata in un notebook Python open source, come quello di Google Colab, un servizio Jupyter Notebook che non richiede configurazione e offre accesso alle risorse di elaborazione, incluse GPU e TPU, rendendolo adatto per machine learning, data science e formazione.

Il flusso di lavoro si articola in due fasi principali (cfr. Fig. 1): classificazione e rilevazione della perplessità; analisi controfattuale guidata dalla perplessità.

Figura 1. Classificazione e Misurazione. Il commento viene classificato e posizionato in uno spazio n-dimensionale, poi per ogni parola rimossa si misura il suo spostamento in questo spazio

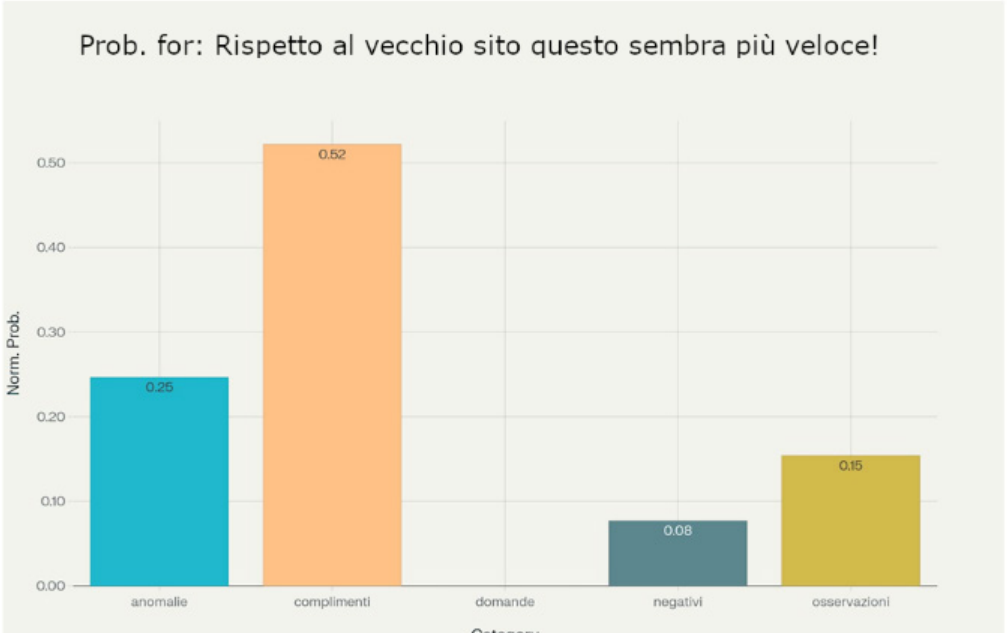


Fonte: Elaborazioni a cura degli Autori

Classificazione token-level con prompt strutturato

Per ogni commento, viene costruito un prompt strutturato che concatena il testo del commento e le possibili categorie di classificazione. Viene utilizzato l'approccio many-shot in-context learning, mostrando al modello un numero ottimale di esempi di classificazione per categoria, tra 125-500 esempi come suggerito da Agarwal *et al*. (2024), utilizzando un prompt standardizzato che richiede semplicemente una classificazione dopo aver fornito gli esempi per ogni categoria. Questo approccio minimizza l'impatto del resto del prompt e del LLM specifico scelto. Il modello genera le probabilità (logprobs) associate ai token di categoria, che vengono estratte per mappare il commento nello spazio vettoriale delle categorie di interesse.

Grafico 1. Probabilità per categoria per il commento "Rispetto al vecchio sito questo sembra più veloce!"



Fonte: Elaborazioni a cura degli Autori

Calcolo della perplessità e configurazioni tecniche

La perplessità è calcolata come inverso della probabilità logaritmica normalizzata assegnata dal modello LLM alla sequenza testuale, data la distribuzione calcolata per le categorie di classificazione. Per i testi modificati (con parole rimosse iterativamente) il calcolo è eseguito identicamente, consentendo la misura della variazione di perplessità come indicatore di importanza lessicale.

Per la riduzione dimensionale e la visualizzazione, è stato utilizzato l'algoritmo t-SNE con i parametri standard consigliati nella letteratura (perplessità=30, learning rate=200, dimensioni output=2 e 3), implementato tramite scikit-learn, per proiettare i vettori probabilistici a 5 dimensioni in uno spazio bi- e tri- dimensionale. Le heatmap per la visualizzazione dell'importanza delle parole sono state generate aggregando le variazioni di perplessità sui singoli termini.

Analisi controfattuale e calcolo della perplessità

Per valutare la rilevanza delle singole parole, si applica una procedura controfattuale: ogni parola del commento viene rimossa iterativamente, il prompt viene ricostruito

e il modello calcola nuovamente la distribuzione di probabilità sulle categorie. La variazione della perplessità tra la classificazione originale e quella modificata funge da indicatore quantitativo dell'importanza della parola rimossa.

Aggregazione e visualizzazione delle spiegazioni

Le variazioni di perplessità vengono aggregate per ciascun commento, evidenziando i termini più influenti. Viene costruita una heatmap dell'importanza relativa delle singole parole. Per visualizzare la posizione di partenza di ogni feedback preso in considerazione viene mappato in un piano bidimensionale, dopo la riduzione della dimensionalità con t-SNE (perplessità=30, learning rate=200, dimensioni output=2 e 3), un algoritmo di riduzione della dimensionalità sviluppato da Geoffrey Everest Hinton e Laurens van der Maaten (2008), utilizzato come strumento di apprendimento automatico in molti ambiti di ricerca.

4.1. Pseudocodice

Mappatura di ciascun commento nello spazio vettoriale delle categorie

```
# Input:
# lista_commenti: lista di stringhe, ciascun commento da classificare
# lista_categorie: lista di stringhe, i simboli corrispondenti alle categorie
# modello_LLM: istanza del modello LLM
# prompt_istruzioni: istruzioni all'LLM su come eseguire il compito
# esempi_manyshot: esempi many-shot per facilitare il fine-tuning del prompt

# Output:
# lista_vettori_probabilita: lista di dizionari con la probabilità normalizzata per ogni categoria per ciascun commento

lista_vettori_probabilita = [ ]
for commento in lista_commenti:
    vettore_prob = [ ]
    for categoria in lista_categorie:
# Costruisci il prompt strutturato
    prompt = prompt_istruzioni + esempi_manyshot + commento
# Invocazione del modello LLM con il prompt per ottenere la distribuzione logprob del token successivo
logprobs_token = modello_LLM.logprob_token(prompt, lista_categorie)
# Il risultato è un dizionario {categoria: logprob_token_corrispondente}
```

```
    # Se il modello non restituisce logprob per qualche categoria,
    assegnare -inf per la normalizzazione
        for cat in lista_categorie:
            if cat not in logprobs_token:
                logprobs_token[cat] = float('-inf')

    # Conversione dei logaritmi di probabilità in probabilità tramite softmax
    max_logprob = max(logprobs_token.values())
    exp_scores = {cat: np.exp(logprob - max_logprob) for cat, logprob in logprobs_
    token.items()}
    somma_exp = sum(exp_scores.values())
    vettore_prob = {cat: exp_scores[cat]/somma_exp for cat in lista_categorie}

    # Aggiungi all'output il vettore di probabilità per il commento corrente
    lista_vettori_probabilita.append(vettore_prob)

    # Fine ciclo

    # lista_vettori_probabilita ora contiene per ogni commento un dizionario con la pro-
    babilità normalizzata per ogni categoria
    # Questi vettori possono essere usati per la mappatura nello spazio a len(lista_cate-
    gorie) dimensioni
```

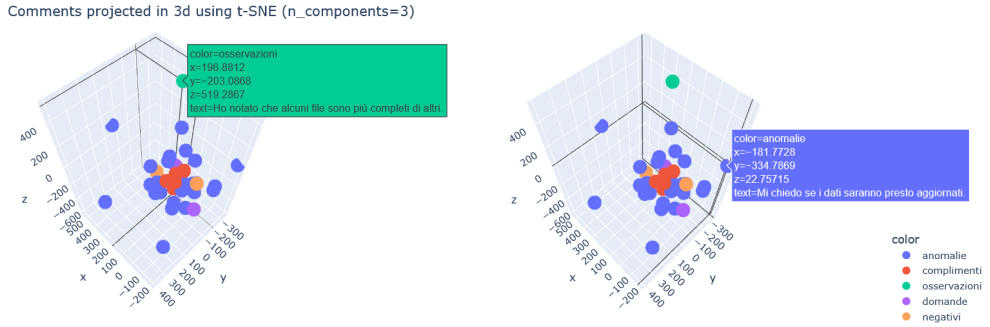
Rimozione iterativa delle parole e valutazione del cambiamento di perplessità

```
    # Input:
    # commento: stringa, testo completo del commento da analizzare
    # lista_categorie: lista di simboli per ogni categoria di classificazione
    # modello_LLM: istanza del modello LLM con funzione per calcolare la perplessità
associata a un testo e alle categorie

    # Output:
    # lista_importanza_parole: lista di float, ciascun valore rappresenta la variazione di
perplessità causata dalla rimozione della parola corrispondente

    # Step 1: suddividi il commento in parole
    parole = commento.split()
```

```
# Step 2: calcola la perplessità del commento originale rispetto al modello e alle categorie
perplessita_originale = calcola_perplessita(commento, lista_categorie, modello_LLM)

lista_importanza_parole = [ ]

# Step 3: iterazione sulle singole parole del commento
for i in range(len(parole)):
# Step 3.1: rimuovi la parola i-esima creando un nuovo testo modificato
commento_modificato = " ".join(parole[:i] + parole[i+1:])

# Step 3.2: calcola la perplessità del testo modificato con la parola rimossa
perplessita_modificata = calcola_perplessita(commento_modificato, lista_categorie, modello_LLM)

# Step 3.3: calcola la variazione di perplessità come differenza tra il testo modificato e il testo originale
delta_perplessita = perplessita_modificata - perplessita_originale

# Step 3.4: aggiungi la variazione alla lista delle importanze
lista_importanza_parole.append(delta_perplessita)

# Step 4: restituisci la lista di importanza delle singole parole
# Valori maggiori di delta_perplessita indicano che la rimozione della parola ha aumentato la perplessità,
# quindi quella parola è più importante nella determinazione della classificazione dal modello.
```

5. Risultati Sperimentali

Grafico 2. Lo spazio 5-dimensionale in 3D. Si nota chiaramente il cluster dei commenti classificati come "complimenti"



Fonte: Elaborazioni a cura degli Autori

I risultati sperimentali ottenuti utilizzando DeepSeek V3, presenti nella tabella 2, dimostrano l'efficacia della metodologia proposta sia nella classificazione sia nell'interpretazione dei feedback degli utenti. L'analisi della distribuzione delle probabilità sui token ha permesso di ottenere una classificazione accurata, con una precisione media del 76% e un F1-score del 74% sulle cinque categorie predefinite, in linea con quanto riportato da Huang e Wang (2025) per un modello della stessa famiglia, DeepSeek R1. In particolare, le categorie "anomalie" e "complimenti" hanno mostrato le performance migliori, con un F1-score rispettivamente dell'82% e dell'80%, mentre le categorie "domande" e "osservazioni" hanno precisione minore, attestandosi intorno al 68% e al 70%.

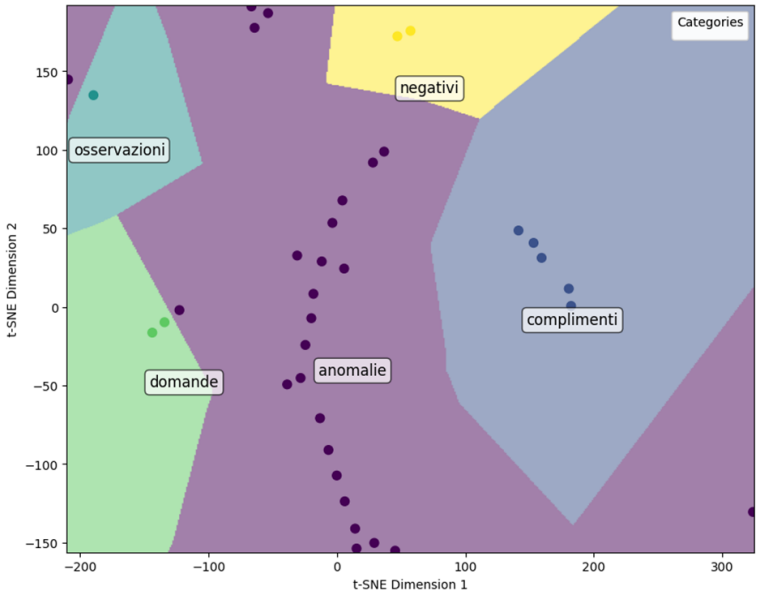
Tabella 2. Categoria/Accuratezza

Categoria	Accuratezza, misurata in F1-score (%)
Anomalie	82
Complimenti	80
Negativi	75
Osservazioni	70
Domande	68

Fonte: Elaborazioni a cura degli Autori

Questo risultato suggerisce che il nostro metodo di Generation as Classification è particolarmente efficace nel distinguere tra feedback positivi e negativi o problematici, ma può richiedere un affinamento ulteriore per cogliere le sfumature tra richieste di informazioni e osservazioni generiche.

*Grafico 3. Mappa bidimensionale dei commenti**

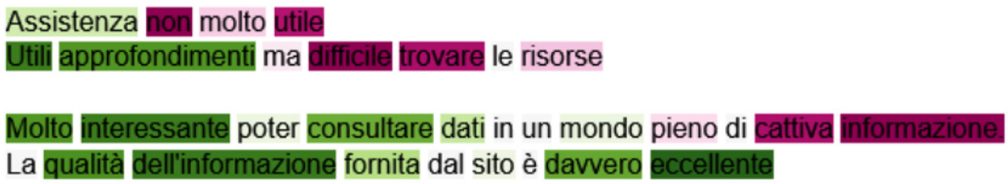


Fonte: Elaborazioni a cura degli Autori

** In base a dove l'LLM posiziona ogni commento possiamo notare se è una classificazione più o meno certa. Si nota chiaramente il cluster delle "anomalie" e quello dei "complimenti", le due categorie meglio identificate dal modello, mentre le "domande" e le "osservazioni" sono più vicine al confine di categoria, il ché fornisce una spiegazione visuale al perché il modello le abbia classificate con meno precisione.*

L'analisi controfattuale guidata dalla perplessità ha rivelato che la rimozione di parole chiave specifiche, come aggettivi qualificativi o nomi propri, ha causato un aumento significativo della perplessità del modello, indicando la loro elevata rilevanza nel processo decisionale. Al contrario, la rimozione di articoli o preposizioni ha avuto un impatto trascurabile sulla perplessità, confermando la capacità del metodo di identificare i termini semanticamente più influenti. La visualizzazione tramite heatmap ha chiaramente evidenziato queste correlazioni, fornendo una rappresentazione intuitiva dell'importanza delle parole.

Figura 2. Heatmap dello spostamento di un commento lungo l'asse "complimenti" - "negativi" dopo la rimozione di ogni parola. Questo consente di avere una spiegazione visuale del contributo di ogni parola alla classificazione



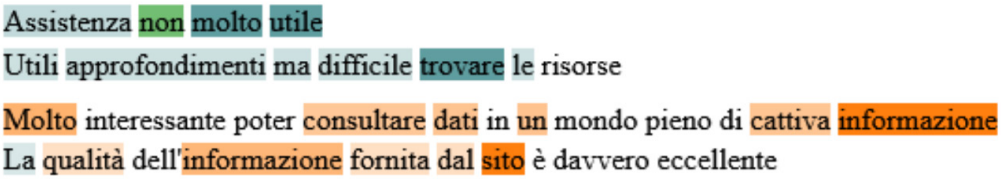
Fonte: Elaborazioni a cura degli Autori

La riduzione della dimensionalità con t-SNE (perplessità=30, learning rate=200, dimensioni output=2 e 3) ha permesso di visualizzare la distribuzione dei feedback nello spazio semantico delle categorie. I cluster formati dai commenti hanno mostrato una buona separazione tra le diverse categorie. Questo risultato conferma che la rappresentazione probabilistica continua a catturare efficacemente le relazioni semantiche tra i commenti e le categorie, facilitando l'identificazione di pattern e anomalie. In particolare, è stato osservato che i commenti classificati erroneamente tendevano a posizionarsi ai confini tra i cluster, suggerendo che l'incertezza del modello è maggiore per questi casi.

5.1. Confronto quantitativo con LIME e SHAP

Per valutare empiricamente l'efficacia della metodologia proposta rispetto a metodi consolidati di interpretazione, è stata condotta un'analisi comparativa su un sottoinsieme del dataset (n=250 commenti). LIME e SHAP sono stati applicati per stimare l'importanza delle parole relative alle stesse categorie testuali. Date le limitazioni di questi nell'applicazione diretta ai modelli linguistici sequenziali, abbiamo implementato un modello di baseline basato su regressione logistica (logreg) con rappresentazione bag-of-words del testo, al quale sono stati applicati LIME e SHAP per l'estrazione dell'importanza delle parole. Questo approccio, pur rappresentando una semplificazione rispetto alla complessità degli LLM, fornisce un punto di confronto metodologico rilevante per valutare le diverse strategie di attribuzione dell'importanza lessicale. Tuttavia si evidenziano anche le limitazioni che i due metodi citati, almeno nelle loro versioni tradizionali, soffrono quando applicati ad approcci complessi al testo quali gli LLM.

*Figura 3. Heatmap dello spostamento di un commento lungo l'asse "complimenti" - "negativi" secondo LIME**



Fonte: Elaborazioni a cura degli Autori

** cfr. Fig. 2 per il confronto con i risultati del nostro metodo.*

Tabella 3. Confronto delle preferenze dei redattori fra LIME, SHAP e il nostro metodo

LIME	SHAP	Ours
10%	10%	80%

Fonte: Elaborazioni a cura degli Autori

In una indagine interna alla redazione dell'ISTAT (cfr. Tab. 3), quattro redattori su cinque hanno ritenuto più realistiche le spiegazioni fornite dal nostro modello (cfr. Graf. 3) rispetto a quelle fornite da LIME (cfr. Graf. 4) o da SHAP (cfr. Graf. 5). Per quanto il sample size sia piccolo, i risultati mostrano che il nostro approccio basato su logprobs e analisi controfattuale di perplessità fornisce spiegazioni più coerenti e localmente precise rispetto a LIME e SHAP, che si basano su un sistema di classificazione meno accurata.

Tabella 4. Confronto Categoria/Accuratezza fra la logreg alla base di LIME e SHAP e il nostro metodo

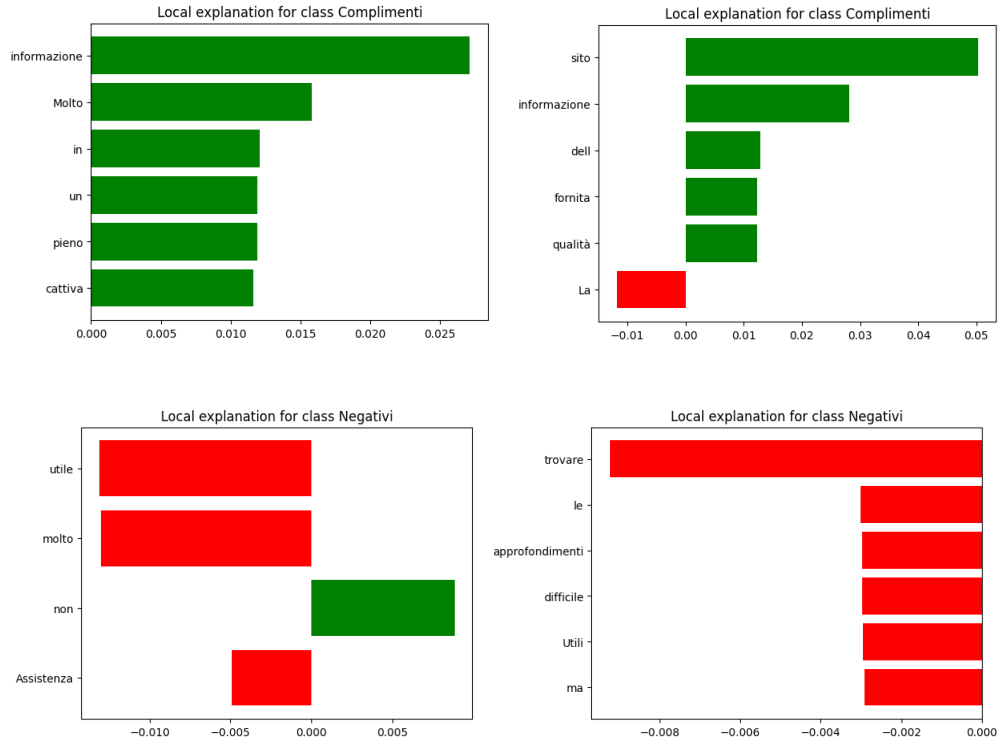
Categoria	Accuratezza LLM, F1-score (%)	Accuratezza logreg, F1-score (%)
Anomalie	**82**	57
Complimenti	**80**	56
Negativi	75	**76**
Osservazioni	**70**	43
Domande	**68**	65

Fonte: Elaborazioni a cura degli Autori

Nel confronto dell'accuratezza nella classificazione per categoria, la classificazione generata da LLM (Deepseek v3) si è dimostrata molto più affidabile della regressione logistica che rappresenta il benchmark di riferimento.

Va considerato che al contrario del nostro metodo, LIME o SHAP sono stati implementati utilizzano questa semplice regressione logistica. Il limite della regressione logistica è che non rende possibile valutare la semantica data dalla sequenza delle parole nel testo libero. La frase "molto chiaro non trovo difficoltà" per la regressione logistica è identica alla frase "non molto chiaro trovo difficoltà", con il metodo proposto invece questa problematica viene superata identificando correttamente la sequenza delle singole parole presenti in un commento libero.

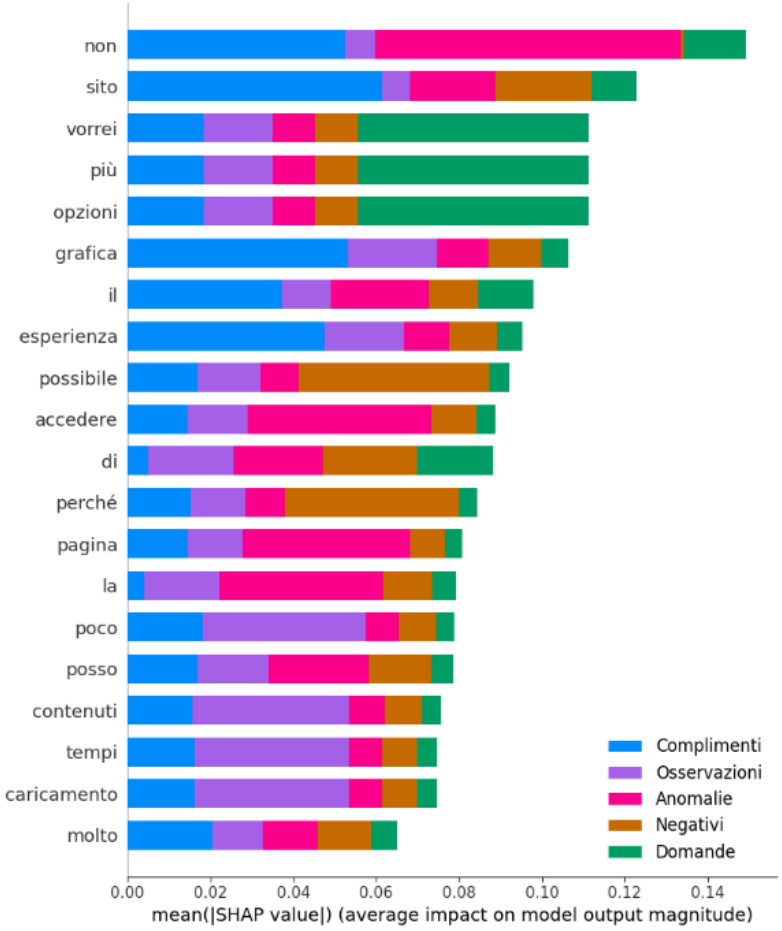
Grafico 4. Spiegazioni locali di LIME per i commenti di Figura 2 e Figura 3



Fonte: Elaborazioni a cura degli Autori

SHAP oltre a fornire spiegazioni a livello di singolo commento, consente anche di aggregare queste informazioni per comprendere globalmente come il modello prende decisioni su tutti i commenti. Per esempio, SHAP può mostrare quali parole sono generalmente più importanti per la classificazione sull'insieme dei commenti e come queste influenzano il comportamento complessivo del modello.

Grafico 5. Valori globali SHAP per le parole più significative nel dataset



Fonte: Elaborazioni a cura degli Autori

In futuro prevediamo di segnalare a editor umani i commenti più vicini ai confini fra le categorie, dato che in questi è più probabile che il modello abbia generato una classificazione errata.

6. Discussione

La metodologia proposta offre un approccio robusto e interpretabile per la classificazione automatica dei feedback testuali, superando alcune delle limitazioni dei metodi tradizionali di XAI applicati ai LLM. L'integrazione della classificazione token-level con l'analisi controfattuale guidata dalla perplessità fornisce non solo una predizione accurata, ma anche una spiegazione granulare delle decisioni del modello, identificando le parole più influenti nel contesto del feedback. Questo è particolarmente rilevante per le statistiche ufficiali, dove la trasparenza e la fiducia nei sistemi automatizzati sono fondamentali.

Il confronto con LIME e SHAP ha evidenziato che, sebbene questi ultimi siano strumenti potenti per l'interpretazione dei modelli, la loro applicazione diretta ai LLM può presentare sfide legate alla natura non strutturata del testo e alla complessità computazionale. Il nostro approccio, progettato specificamente per l'analisi testuale e lavorando direttamente sulle probabilità dei token e sulla perplessità, si adatta meglio alle specificità dei LLM, offrendo un meccanismo di spiegazione più nativo e meno oneroso. La capacità di generare spiegazioni locali e di visualizzare l'importanza delle parole tramite heatmap rende il sistema accessibile anche a utenti non esperti di IA, facilitando la comprensione e la validazione delle classificazioni.

Un aspetto cruciale emerso dalla letteratura è la diversità delle spiegazioni controfattuali. Mentre la nostra metodologia si concentra sulla rilevanza delle singole parole, futuri sviluppi potrebbero esplorare la generazione di controfattuali più complessi, che coinvolgano modifiche a intere frasi o al contesto semantico, per offrire una gamma più ampia di alternative interpretabili. L'utilizzo di un prompt standardizzato e dell'approccio many-shot in-context learning minimizza l'impatto della formulazione specifica del prompt e del LLM scelto, garantendo maggiore robustezza e coerenza delle classificazioni anche in presenza di variazioni linguistiche o di ambiguità nel feedback.

7. Conclusioni

In questo lavoro abbiamo presentato una metodologia innovativa per la classificazione e l'interpretazione dei feedback testuali tramite LLM, integrando l'analisi delle probabilità token-level con tecniche controfattuali guidate dalla perplessità.

L'approccio proposto consente di ottenere classificazioni accurate e di identificare le parole più rilevanti per le decisioni del modello, migliorando la trasparenza e l'analizzabilità dei sistemi di IA. I risultati sperimentali hanno dimostrato l'efficacia del metodo, evidenziando la sua capacità di fornire spiegazioni granulari e intuitive attraverso l'utilizzo di tecniche many-shot in-context learning che superano le limitazioni degli approcci tradizionali few-shot.

8. Futuri Sviluppi

Il presente lavoro apre diverse direzioni per futuri sviluppi. In primo luogo, l'esplorazione di tecniche avanzate di generazione di controfattuali, che considerino non solo la rimozione di singole parole ma anche la sostituzione o l'aggiunta di intere frasi, potrebbe arricchire la diversità e la completezza delle spiegazioni. Questo includerebbe l'adozione di metriche di diversità più sofisticate, come quelle discusse da Laugel *et al*. (2023), per garantire che le spiegazioni controfattuali offrano una gamma completa di scenari alternativi.

In secondo luogo, l'integrazione di meccanismi per la gestione dell'incertezza del modello, come suggerito da Thuy e Benoit (2024), potrebbe migliorare la robustezza delle classificazioni e delle spiegazioni, fornendo agli utenti una misura più affidabile della confidenza del modello. Questo potrebbe tradursi in un sistema che non solo classifica i feedback, ma indica anche quando la sua predizione è meno certa, permettendo un intervento umano mirato.

Un'area di sviluppo futuro riguarda il perfezionamento del calcolo della perplessità. L'attuale implementazione utilizza un approccio semplificato che potrebbe non tenere conto delle complessità delle architetture LLM moderne. Sviluppi futuri potrebbero esplorare metodi più sofisticati per il calcolo della perplessità che considerino le relazioni più complesse tra token-level probabilities e confidenza del modello, incorporando aspetti come la calibrazione delle probabilità e la gestione dell'incertezza epistemica e aleatoria.

Infine, l'applicazione della metodologia a dataset di feedback più ampi e diversificati, provenienti da contesti diversi dalle statistiche ufficiali, consentirebbe di valutarne la generalizzabilità e di identificare eventuali adattamenti necessari. L'obiettivo è sviluppare un framework di XAI per i LLM che sia versatile, efficiente e in grado di supportare una vasta gamma di applicazioni, contribuendo a costruire sistemi di intelligenza artificiale più trasparenti e affidabili.

Bibliografia

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., & Larochelle, H. (2024). Many-shot in-context learning. *arXiv preprint arXiv:2404.11018* https://doi.org/10.48550/arXiv.2404.11018.

Chatterjee, S, Colombo, E., and Raimundo, M., Multi-criteria Rank-based Aggregation for Explainable AI, In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE. https://doi.org/10.1109/IJCNN64981.2025.11228222.

Cui, Y., He, P., Zeng, J., Liu, H., Tang, X., Dai, Z., Han, Y., Luo, C., Huang, J., Li, Z., Wang,

S., Xing, Y., Tang, J., & He, Q. (2025). Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2502.13260* https://doi.org/10.48550/arXiv.2502.13260.

De Oliveira, R. M. B., & Martens, D. (2021). A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences,* 11(16), 7274. https://doi.org/10.3390/app11167274.

Huang, D., & Wang, Z. (2025). Explainable sentiment analysis with DeepSeek-R1: Performance, Efficiency, and Few-Shot Learning. *arXiv preprint arXiv:2503.11655* https://doi.org/10.48550/arXiv.2503.11655.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. (1977). Perplexity - a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1), S63–S63. https://doi.org/10.1121/1.2016299.

Laugel, T., Jeyasothy, A., Lesot, M.-J., Marsala, C., & Detyniecki, M. (2023). Achieving diversity in counterfactual explanations: A Review and Discussion. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. https://doi.org/10.1145/3593013.3594122.

Liu, J., Wu, X., Liu, S., & Gong, S. (2025). Model-agnostic counterfactual explanation: A feature weights-based comprehensive causal multi-objective counterfactual framework. *Expert Systems with Applications*, 260, Article 126063. https://doi.org/10.1016/j.eswa.2024.126063.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)* (pp. 607-617). https://doi.org/10.1145/3351095.337285.

Samoilescu, R. F., Van Looveren, A., & Klaise, J. (2021). Model-agnostic and scalable counterfactual explanations via reinforcement learning. *arXiv preprint arXiv:2106.02597*. https://doi.org/10.48550/arXiv.2106.02597.

Stepin, I., Alonso, J. M., Catalá, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9, 11974-12001. https://dx.doi.org/10.1109/ACCESS.2021.3051315.

Thuy, A., & Benoit, D. F. (2024). Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, 312(3), 330-340 https://doi.org/10.1016/j.ejor.2023.09.009.

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.htm.

Yu, Y.-C., Kuo, C.-C., Ye, Z., Chang, Y.-C., & Li, Y.-S. (2024). Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 1826-1839). https://doi.org/10.18653/v1/2024.findings-emnlp.99.

DECOSTRUIRE L'AI: TRA PAURE PUBBLICHE E FONDAMENTI SCIENTIFICI

DECONSTRUCTING AI: BETWEEN PUBLIC FEARS AND SCIENTIFIC FOUNDATIONS

Tonio Di Battista[1]

Sommario

Questo articolo esamina criticamente il discorso contemporaneo sull'intelligenza artificiale (AI), mettendo in discussione le concezioni errate più diffuse e sottolineando la necessità di adottare una prospettiva scientificamente fondata. Sostiene che i sistemi di AI attuali, basati principalmente su modelli statistici e ottimizzazione computazionale, non possiedono una cognizione autonoma, contrariamente alle narrazioni popolari che tendono ad antropomorfizzare l'AI attribuendole coscienza o creatività umane.

L'AI, piuttosto che replicare l'intelligenza umana, eccelle nell'elaborazione di grandi quantità di dati attraverso l'inferenza probabilistica, supportando il processo decisionale in ambiti complessi come la sanità, la finanza e l'istruzione. L'articolo evidenzia come i metodi moderni di AI – apprendimento automatico, apprendimento profondo e reti neurali – siano estensioni della statistica classica, con i progressi recenti dovuti più alla potenza computazionale che a innovazioni concettuali.

Il termine "intelligenza" viene criticato per alimentare confusione semantica e timori infondati riguardo al potenziale dell'AI di superare o sostituire le capacità umane. L'autore promuove una maggiore alfabetizzazione tecnica e una collaborazione interdisciplinare, invitando decisori politici e ricercatori a basare la governance dell'AI su una chiarezza metodologica piuttosto che su etiche speculative.

Abstract

This paper critically examines contemporary discourse on artificial intelligence (AI), challenging prevalent misconceptions and highlighting the need for a scientifically grounded perspective. It argues that current AI systems, primarily based on statistical modeling and computational optimization, lack autonomous cognition, contrary to popular narratives that anthropomorphize AI with human-like consciousness or creativity.

[1] Università degli studi "G. d'Annunzio", Dipartimento di Studi Socio-Economici, Gestionali e Statistici, Chieti-Pescara, Italia - e-mail: tonio.dibattista@unich.it

Rather than replicating human intelligence, AI excels in processing large datasets via probabilistic inference, supporting decision-making in complex domains such as healthcare, finance, and education. The article emphasizes that modern AI methods – machine learning, deep learning, and neural networks – are extensions of classical statistics, with recent progress driven more by computational power than by conceptual breakthroughs.

The term "intelligence" is critiqued for fostering semantic confusion and unwarranted fears about AI's potential to surpass or replace human capabilities. It promotes greater technical literacy and interdisciplinary collaboration, urging policymakers and researchers to base AI governance on methodological clarity rather than speculative ethics.

Parole chiave: Intelligenza artificiale, etica, statistica, ricerca scientifica.
Keywords: *Artificial intelligence, ethics, statistics, scientific research.*

Incipit:
"L'intelligenza artificiale – nella sua attuale configurazione – non è una forma di intelligenza autonoma, ma un "pappagallo stocastico", è potente ma privo di consapevolezza. Spetta all'uomo – e in particolare allo statistico – il compito di guidare l'uso di queste tecnologie, evitando derive irrazionali o automatismi pericolosi."

GIORGIO PARISI

1. Introduzione

Negli ultimi anni, l'intelligenza artificiale (AI) è diventata uno dei temi più discussi nel dibattito pubblico, suscitando entusiasmo, timori e interrogativi profondi di natura etica, filosofica e sociale. La sua rapida diffusione in ambiti strategici come la medicina, la finanza, l'industria e l'educazione ha alimentato l'idea che queste tecnologie possano, in un futuro prossimo, eguagliare o addirittura superare l'intelligenza umana. Tuttavia, tale percezione è spesso il frutto di una comprensione distorta delle reali capacità dell'AI, amplificata da una narrazione mediatica imprecisa e da un linguaggio che tende ad antropomorfizzare strumenti puramente matematici e statistici.

L'obiettivo di questo contributo è quello di chiarire, su base scientifica, la natura e i limiti strutturali dell'intelligenza artificiale nella sua configurazione attuale. Si intende dimostrare come l'AI moderna, lungi dall'essere una "mente artificiale" autonoma, rappresenti in realtà un'evoluzione scalabile della modellistica statistica tradizionale, resa possibile dall'aumento della potenza computazionale e dalla disponibilità di grandi quantità di dati. I modelli di machine learning e deep learning e transfer learning, oggi

al centro delle applicazioni più avanzate, non sono altro che algoritmi ottimizzati per compiti specifici, privi di coscienza, intenzionalità o comprensione del contesto.

In questa prospettiva, il dibattito sull'AI dovrebbe spostarsi da un piano immaginifico e spesso allarmistico a uno più concreto e metodologicamente fondato, che valorizzi le competenze scientifiche e promuova un approccio interdisciplinare. Solo attraverso una maggiore alfabetizzazione tecnica e una riflessione critica basata su dati e modelli rigorosi sarà possibile affrontare le reali sfide poste da queste tecnologie, evitando derive ideologiche e costruendo un dialogo equilibrato tra sapere umano e strumenti artificiali.

2. Intelligenza artificiale: opinioni e fatti

Il tema dell'intelligenza artificiale (AI) è oggetto di un dibattito pubblico sempre più acceso, al centro del quale vi sono interrogativi etici, filosofici e perfino ontologici. Al crescere delle sue applicazioni in ambiti come la medicina, la finanza, l'industria, la didattica e l'informazione, si è generato un clima di entusiasmo ma anche di allarme: si paventa una possibile sostituzione dell'intelligenza umana da parte di sistemi automatizzati, capaci di apprendere, agire e persino "decidere" (Haenlein and Kaplan, 2019).

È opportuno sottolineare che un sistema di intelligenza artificiale può non solo fornire risposte, ma anche formulare domande. Tuttavia, è importante precisare che la capacità dell'AI di porre domande non implica un'autentica curiosità né un'intenzionalità cognitiva. Le domande generate derivano infatti dalle informazioni e dai pattern presenti nei dati su cui il sistema è stato addestrato.

In altri termini, l'AI può formulare quesiti pertinenti all'ambito delle conoscenze già acquisite, ma non può interrogarsi su concetti o oggetti che non rientrano nel suo spazio informativo, poiché gli strumenti di deep learning, machine learning e i Large Language Models (LLM) si basano su dati preesistenti. Anche i processi di intelligenza generativa risultano quindi vincolati alle conoscenze acquisite e non possono produrre concetti totalmente estranei all'esperienza o ai dati disponibili.

Questa narrazione, tuttavia, appare spesso distorta da una lettura superficialmente antropomorfica della tecnologia, alimentata da un linguaggio impreciso e da una diffusa carenza di comprensione tecnica. Si discute di "coscienza artificiale", intesa come l'ipotetica capacità di un sistema artificiale di possedere un'esperienza soggettiva o una consapevolezza di sé, un concetto che al momento resta più filosofico che scientificamente fondato. Si parla poi di "macchine senzienti", cioè entità tecnologiche in grado di provare sensazioni, emozioni o stati mentali analoghi a quelli umani, un'idea ancora priva di riscontro empirico. Si discute quindi di "super intelligenze" in grado di sfuggire al controllo umano, ma raramente tali considerazioni sono accompagnate da un'analisi approfondita dei presupposti scientifici e computazionali dell'AI.

Un aspetto particolarmente problematico del dibattito sull'AI risiede nel fatto che i principali promotori della discussione pubblica – filosofi morali, intellettuali, giornalisti – sono spesso estranei alle discipline tecnico-scientifiche che sottendono lo sviluppo degli algoritmi di apprendimento automatico. La loro riflessione, seppur culturalmente legittima, tende a ruotare attorno a questioni astratte di ordine morale, ontologico o sociale, e a trascurare le basi empiriche e matematiche che definiscono concretamente cosa l'intelligenza artificiale possa o non possa fare. Questo squilibrio produce un'eccessiva enfasi sul rischio che tali tecnologie possano sostituire – o peggio ancora, sovrastare – l'essere umano, senza comprendere che l'AI, nella sua forma attuale, non possiede alcuna intenzionalità autonoma, ma agisce esclusivamente in base a schemi probabilistici e pattern estratti da dati pregressi.

Il problema centrale sembra dunque risiedere in una profonda lacuna culturale circa il funzionamento intrinseco dei sistemi di intelligenza artificiale. In pochi, anche tra i decisori politici e i professionisti dell'informazione, sanno realmente cosa siano il *machine learning*, il *deep learning* (Lecun *et al*., 2015, Tosi, 2020), o le tecniche di *classificazione supervisionata* e *non supervisionata* (MacKay, 2005). Ancora meno diffusa è la consapevolezza che tali metodi siano il frutto di una lunga evoluzione della statistica computazionale e dell'analisi matematica, basati su logiche inferenziali rigorose. Si tratta di strumenti di modellizzazione che apprendono a partire da esempi forniti tramite grandi dataset, ma che non hanno capacità autonome di astrazione, intuizione o giudizio etico. Questa ignoranza tecnica alimenta una narrazione mistificata dell'AI, che ne sovrastima le capacità reali e ne sottovaluta al contempo i limiti strutturali.

3. L'AI come evoluzione della modellistica statistica: una trasformazione di scala, non di principio

L'avvento dell'intelligenza artificiale rappresenta una trasformazione profonda e strutturale nel trattamento dei dati, più che una rivoluzione dei modelli concettuali. Ciò che oggi definiamo "AI moderna" si fonda in larga parte su estensioni avanzate della modellistica statistica classica, arricchite da tecnologie computazionali di nuova generazione. Le tecniche di machine learning (apprendimento automatico), deep learning (apprendimento profondo) e reti neurali più in generale, che costituiscono oggi il cuore operativo dell'intelligenza artificiale, derivano in modo diretto da principi teorici elaborati nella statistica e nell'analisi matematica dei dati fin dal secolo scorso (Chen *et al*., 1996).

Molti algoritmi oggi alla base delle applicazioni industriali e scientifiche più avanzate – come la regressione logistica per la classificazione binaria, i modelli bayesiani per l'analisi inferenziale, le reti neurali artificiali per la modellazione non lineare, e le tecniche di clustering per l'esplorazione non supervisionata – affondano le proprie radici

nella teoria statistica consolidata. Ad esempio, il perceptrone (Bishop, 1994), precursore delle moderne reti neurali, fu introdotto negli anni '50 con una base matematica affine a quella della regressione lineare multivariata. I modelli bayesiani, d'altra parte, poggiano su fondamenti probabilistici risalenti a Thomas Bayes (XVIII secolo) (Chickering, 1996), oggi implementati a larga scala mediante tecniche computazionali come il Monte Carlo Markov Chain (MCMC).

La vera discontinuità introdotta dall'intelligenza artificiale non risiede nella natura teorica dei modelli, ma nella scala su cui questi modelli operano. Le tecniche di machine learning sono state rese praticabili solo grazie all'incremento esponenziale della potenza computazionale e alla disponibilità di dataset massivi, raccolti tramite sensori, dispositivi digitali e piattaforme online. Si tratta di un cambiamento quantitativo che ha prodotto effetti qualitativi: i modelli possono ora essere addestrati su miliardi di parametri (si pensi ai LLM – Large Language Models), e sono capaci di affinare iterativamente le proprie previsioni, correggendo gli errori in base al feedback e migliorando nel tempo attraverso processi di ottimizzazione come il gradient descent.

Un esempio emblematico è rappresentato dal deep learning, che utilizza reti neurali profonde (composte da decine o centinaia di livelli) per modellare relazioni complesse e non lineari nei dati, come il riconoscimento di immagini, la sintesi vocale o la generazione linguistica. Queste architetture, formalmente complesse ma teoricamente ancora basate su funzioni di attivazione e pesatura degli input, sono state rese possibili solo grazie alla capacità di elaborare milioni di esempi in tempi compatibili con l'uso applicativo.

Nonostante l'evoluzione tecnica, la funzione epistemologica della statistica rimane immutata. L'obiettivo ultimo resta quello di comprendere i fenomeni osservabili, validare le ipotesi interpretative e fornire inferenze affidabili a partire da dati incompleti, rumorosi o incerti. La statistica si occupa da sempre di modellare la realtà in modo probabilistico, di stimare parametri ignoti e di valutare l'incertezza associata a ciascuna conclusione. In questo senso, l'AI moderna rappresenta un potenziamento di questa funzione.

Anche nei sistemi più sofisticati, come i modelli generativi basati su transformer (es. GPT, BERT), i principi statistici sono alla base della costruzione delle distribuzioni di probabilità condizionate, che guidano la generazione linguistica o la classificazione semantica. La differenza risiede nella complessità e nella dimensionalità dello spazio dei dati, ma il fondamento matematico e statistico resta centrale. In breve, l'intelligenza artificiale si fonda sul metodo statistico, rendendolo scalabile, flessibile e applicabile a problemi prima inaffrontabili per limiti computazionali. La relazione cruciale risiede proprio nel rapporto tra il metodo statistico – che nel tempo ha seguito propri percorsi di sviluppo teorico – e i sistemi di calcolo, i quali evolvono secondo logiche differenti e con una rapidità tecnica di gran lunga superiore rispetto ai tempi della statistica accademica.

4. La confusione semantica: il termine “intelligenza” come fattore fuorviante

Un elemento particolarmente critico nell’alimentare tale confusione è di natura linguistica: il termine “intelligenza”, usato nella locuzione “intelligenza artificiale”, evoca inevitabilmente connotazioni umane, cognitive, morali. L’antropomorfizzazione del concetto induce molti osservatori a proiettare sugli algoritmi caratteristiche che essi non possiedono affatto. In realtà, l’AI non è “intelligente” nel senso umano del termine: non ha intenzionalità, non comprende il mondo in modo simbolico, non ha coscienza di sé, né capacità di introspezione. Essa non fa altro che analizzare dati sulla base di funzioni matematiche ottimizzate per obiettivi specifici (ad esempio, minimizzare l’errore di previsione). Paradossalmente, è proprio l’uso inopportuno di questo termine che ha generato un allarmismo sociale che trascende i reali pericoli legati all’uso irresponsabile o distorto della tecnologia.

Alla base dell’AI vi è dunque un impianto rigorosamente scientifico, derivato dalla statistica, dall’informatica e dalla teoria dell’informazione. Le reti neurali artificiali, ad esempio, non sono che modelli matematici ispirati – solo superficialmente – al funzionamento del cervello umano, progettati per ottimizzare compiti specifici come la classificazione di immagini o la generazione di testo. Il machine learning consiste nell’addestrare un modello su un insieme di dati in modo che possa effettuare previsioni su nuovi casi. Non vi è nulla di “magico” o “misterioso” in questi processi: si tratta di operazioni matematiche, spesso estremamente sofisticate, ma prive di qualsiasi forma di autonomia morale o cognitiva. La percezione diffusa che l’AI possa “sfuggire al controllo” è quindi più il frutto di una distorsione immaginifica che di un rischio concreto. Il vero pericolo, semmai, consiste nell’utilizzo inconsapevole o opaco di questi strumenti, senza adeguate conoscenze metodologiche e senza un chiaro quadro normativo.

Per affrontare in modo maturo e responsabile il tema dell’intelligenza artificiale, è necessario promuovere una cultura tecnica ampia e diffusa. Solo una maggiore alfabetizzazione scientifica e statistica, a partire dalle scuole e fino ai livelli più alti della formazione accademica, può permettere una comprensione corretta dei reali meccanismi che regolano l’AI e dei limiti entro cui essa può operare. Le preoccupazioni etiche non devono essere accantonate, ma vanno integrate con una solida base di conoscenze tecniche (Floridi, 2022). In ultima analisi, la riflessione sull’AI non può prescindere da un approccio interdisciplinare che includa, accanto alla filosofia e all’etica, anche la statistica, l’informatica e la scienza dei dati. Solo così sarà possibile distinguere tra i timori infondati e le sfide concrete, e costruire un rapporto equilibrato tra uomo e macchina, nel rispetto della dignità umana e nella consapevolezza delle possibilità – ma anche dei limiti – della tecnologia.

5. L'intelligenza artificiale come sistema di elaborazione della conoscenza esistente

È ormai evidente che i sistemi di intelligenza artificiale si basano sull'elaborazione strutturata di dati e informazioni già disponibili, attingendo a vaste banche dati e operando mediante modelli statistici complessi. L'AI, nella sua forma attuale, è una tecnologia orientata alla rielaborazione della conoscenza preesistente, non alla creazione di sapere ex novo. Essa funziona tramite due principi fondamentali del metodo statistico: l'inferenza e l'ottimizzazione, nel senso che, attraverso questi processi, i modelli apprendono schemi ricorrenti nei dati e cercano di minimizzare l'errore di previsione. Ma se i dati a disposizione non racchiudono conoscenze avanzate, è impossibile per un sistema di AI generare innovazione autentica o "scoprire" soluzioni tecnologiche mai pensate prima. Questo evidenzia un limite strutturale della tecnologia: l'AI non inventa, ma combina e generalizza informazioni note.

Per comprendere a fondo questa limitazione, basta ricorrere a un esperimento mentale. Se fornissimo a un algoritmo di intelligenza artificiale solo i dati scientifici e tecnologici in possesso dell'umanità nel 1900, potremmo aspettarci che esso inventi un microprocessore, una rete di comunicazione globale, o un dispositivo quantistico? La risposta è no. I dati del 1900 non contenevano né le scoperte fisiche né gli avanzamenti teorici che hanno reso possibile l'elettronica moderna. Allo stesso modo, se l'AI disponesse solo della conoscenza del XVIII secolo, sarebbe mai in grado di concepire un viaggio spaziale, una missione sulla Luna o il concetto stesso di orbita geostazionaria? Questo esempio rende palese che l'intelligenza artificiale non può generare scoperte fuori dallo spazio della conoscenza disponibile, se non in minima parte mediante inferenze marginali. L'innovazione vera, quella che produce discontinuità epistemologiche, resta prerogativa della ricerca scientifica umana e sperimentale, capace di interrogarsi sull'ignoto e di formulare ipotesi radicalmente nuove. Se, da un lato, l'AI non è uno strumento creativo nel senso stretto, dall'altro essa possiede una straordinaria capacità di sintesi. Può integrare simultaneamente conoscenze provenienti da campi diversi, elaborare modelli predittivi sulla base di milioni di variabili, e offrire supporto decisionale in contesti complessi.

Un esempio illuminante proviene dal campo della medicina. Consideriamo due medici: uno appena laureato, aggiornato sulla letteratura più recente, e uno esperto, che ha affrontato centinaia di casi clinici. A parità di formazione teorica, quale dei due sarebbe in grado di diagnosticare meglio un caso clinico atipico? La risposta naturale è: colui che ha avuto accesso a più esperienze, a più dati, a una maggiore varietà di situazioni. Ora immaginiamo un sistema di AI addestrato su milioni di cartelle cliniche, sintomi, diagnosi e trattamenti. Esso rappresenta, potenzialmente, la concentrazione di tutta l'esperienza medica nota, pronta a essere interrogata per fornire risposte rapide, contestuali e basate su dati. Non è "più intelligente" di un medico, non potrà mai sostituire il

medico ma può offrirgli uno strumento di supporto cognitivo senza precedenti. Il medico rimane, e rimarrà in futuro, una figura insostituibile anche nell'era dell'intelligenza artificiale. Sebbene i sistemi di AI possano offrire supporto avanzato nella raccolta e nell'analisi dei dati clinici, sarà fondamentale che il medico possieda le competenze necessarie per interagire in modo critico ed efficace con tali tecnologie. Non solo dovrà saper interrogare correttamente il sistema di AI, ma anche interpretarne i risultati alla luce del contesto clinico specifico del paziente, integrando le informazioni fornite con il proprio giudizio professionale, l'esperienza e la sensibilità umana.

6. Dal panico etico allo sviluppo tecnico: dove andrebbe indirizzato il dibattito

Alla luce di queste considerazioni, appare chiaro che il timore di un'intelligenza artificiale "sostitutiva" dell'uomo è frutto di una profonda incomprensione del funzionamento reale degli algoritmi. L'AI non minaccia l'intelligenza umana, ma può potenziarla, qualora si comprenda e si governi adeguatamente la sua struttura matematica. Piuttosto che alimentare un dibattito fondato sull'ignoranza tecnica, sarebbe più utile promuovere uno studio approfondito dei modelli statistici che costituiscono la base dei sistemi di intelligenza artificiale, concentrandosi sulla loro ottimizzazione e sullo sviluppo di nuovi metodi di previsione e interpretazione dei risultati delle analisi. Ciò richiede una forte spinta verso la ricerca interdisciplinare, capace di integrare competenze ingegneristiche, informatiche e matematiche, e la costruzione di un corpus teorico solido, sul quale innestare – solo in un secondo momento – una riflessione etica e giuridica adeguata. In questa prospettiva, anche la statistica è chiamata a riflettere criticamente sul proprio ruolo all'interno della scienza. Essa non dovrebbe ridursi a un insieme di tecniche formali, ma riaffermarsi come strumento di soluzione di problemi reali, partendo dai fenomeni osservati piuttosto che dai metodi stessi. Solo in questo modo la statistica potrà evitare il rischio di rimanere astratta e marginale rispetto al machine learning, che, pur condividendo le sue radici metodologiche, si distingue per la capacità di operare direttamente sui dati reali e di fornire risposte operative a questioni concrete.

Per rendere ancora più chiaro lo stato attuale dell'AI, è utile ricorrere a un paragone storico. L'intelligenza artificiale oggi è paragonabile al motore a scoppio nei primi anni del Novecento. All'epoca, anche il motore a combustione interna generava dubbi e paure: avrebbe distrutto l'industria del cavallo? Avrebbe causato incidenti mortali? Avrebbe reso l'uomo schiavo della macchina? Quelle preoccupazioni, pur in parte fondate, non hanno impedito il progresso. Il motore ha permesso uno sviluppo industriale senza precedenti, ha ampliato la mobilità, ha stimolato la nascita di nuovi settori economici, ricchezza e benessere sociale. Ma non è nato perfetto. Ha richiesto decenni di miglioramenti tecnici, nuove scoperte nei materiali, nella termodinamica, nell'elettronica. L'AI

è in una fase analoga: è solo all'inizio di un percorso che la renderà – se adeguatamente coltivata – un motore di progresso sociale, culturale ed economico.

L'intelligenza artificiale (AI) è destinata a diventare sempre più efficiente e innovativa, in stretta correlazione con i progressi della ricerca scientifica. È proprio la scienza, infatti, a rappresentare il vero motore evolutivo di tali tecnologie: le scoperte, le teorie e le metodologie emergenti fungeranno da "carburante" per l'elaborazione e il perfezionamento degli algoritmi di AI. Questi ultimi, a loro volta, diventeranno via via più precisi, robusti e capaci di rispettare principi etici, proprio grazie agli avanzamenti ottenuti nei diversi ambiti della ricerca scientifica, inclusi quelli dell'etica computazionale, delle neuroscienze e delle scienze sociali.

Ne consegue che la responsabilità di interpretare i bisogni emergenti della società e di fornire risposte ai nuovi interrogativi posti dall'evoluzione tecnologica spetta innanzitutto alla scienza, non all'intelligenza artificiale che, al pari delle tecnologie esistenti, avrà il ruolo di coadiuvare e contribuire al suo sviluppo e successo. L'AI deve essere considerata per quello che è: uno strumento estremamente potente e veloce nell'elaborazione dei dati, ma privo di coscienza, intenzionalità e capacità autonoma di discernimento.

In questo contesto, è fondamentale ribadire che il progresso tecnologico non può e non deve sostituire il ruolo centrale dell'uomo e della conoscenza scientifica. Al contrario, sarà la scienza, guidata dalla curiosità e dall'ingegno umano, a mantenere la leadership nel determinare la direzione dello sviluppo della civiltà. Pertanto, anche nell'era dell'intelligenza artificiale, l'uomo – attraverso la scienza – resterà il protagonista e il principale artefice del progresso umano.

A questo principio di fondo si affianca oggi una cornice normativa coerente con tale visione antropocentrica. La Legge 23 settembre 2025, n. 132, ha infatti sancito che lo sviluppo, l'adozione e l'uso dei sistemi di intelligenza artificiale devono avvenire in modo trasparente, responsabile e proporzionato, nel rispetto dei diritti fondamentali, della Costituzione e del diritto dell'Unione europea. L'AI deve essere soggetta a sorveglianza e intervento umano, garantire sicurezza, correttezza, affidabilità, protezione dei dati personali, non discriminazione e riservatezza, non compromettere il metodo democratico né la libertà del dibattito politico, e assicurare cybersicurezza lungo tutto il ciclo di vita dei sistemi, nonché accessibilità universale, inclusa quella delle persone con disabilità.

In sostanza, il principio giuridico cardine è che l'intelligenza artificiale è al servizio dell'uomo, non sostitutiva della sua autonomia decisionale, e deve operare entro un quadro di garanzie etiche, giuridiche e di sicurezza che tutelino i diritti, le libertà e la sovranità democratica.

Tuttavia, nel dibattito attuale a volte si esagera nel definire, spesso prematuramente, cornici normative e limiti etici prima ancora di aver compreso a fondo la natura dello

strumento. Tornando al paragone con l'automobile, sarebbe come scrivere un codice della strada prima di sapere se le auto possono viaggiare, a che velocità, con quale carburante, e con quali sistemi di sicurezza. La regolamentazione è necessaria, ma deve essere conseguente alla conoscenza tecnica, non anticiparla. Altrimenti, il rischio è di frenare un potenziale progresso solo per timore dell'ignoto. Inoltre, è fondamentale distinguere tra l'uso strumentale di una tecnologia e i suoi possibili abusi: il primo va incoraggiato e regolato, il secondo prevenuto e sanzionato, secondo il principio di proporzionalità e tutela dei diritti sancito anche dalla recente normativa nazionale.

Nel corso della storia, l'umanità ha spesso dovuto confrontarsi con l'ambivalenza del progresso tecnologico: strumenti e scoperte che, pur mostrando un enorme potenziale di utilità, hanno anche rivelato la loro capacità distruttiva. Il motore a scoppio, ad esempio, ha alimentato non solo la mobilità civile e lo sviluppo industriale, ma anche i carri armati che hanno attraversato i campi di battaglia delle guerre. Analogamente, la scissione nucleare ha portato con sé l'orrore di Hiroshima e Nagasaki, ma non per questo l'umanità ha rinunciato a sfruttarne i benefici, basti pensare alla produzione energetica, alla medicina nucleare e all'esplorazione spaziale ed al conseguente benessere economico e sociale di queste scoperte

Oggi, di fronte all'ascesa dell'intelligenza artificiale, ci troviamo a un bivio simile. Sarebbe miope e limitante lasciare che la paura dell'ignoto, della inconsapevolezza del reale funzionamento o la mancanza di comprensione tecnica frenino l'evoluzione di una delle tecnologie più promettenti del nostro tempo. Il vero freno non deve essere l'aura dell'ignoranza, ma un costante e rigoroso approfondimento culturale, che coinvolge tutte le discipline in primis la statistica che regolamenta il funzionamento degli algoritmi che sono alla base dell'AI, ma anche le discipline umanistiche e filosofiche. Attraverso un dialogo continuo tra scienza, etica e umanesimo è possibile orientare l'intelligenza artificiale verso finalità realmente benefiche, capaci di contribuire all'equità, al benessere e alla sostenibilità globale. Ma tutto questo è stato, è oggi e sarà il compito della ricerca scientifica nella sua totale complessità

L'intelligenza artificiale, infatti, potrà giocare un ruolo determinante non solo come strumento operativo, ma come catalizzatore di riflessione e trasformazione sociale. Il suo impatto sarà tanto più positivo quanto più sarà guidato da un pensiero critico consapevole, radicato nelle esperienze storiche dell'umanità e capace di riconoscere che il progresso non è fine a sé stesso, ma deve essere continuamente misurato in relazione al bene comune. In questa prospettiva, la cultura, la filosofia e le scienze umane non sono accessori, ma componenti essenziali di ogni progettualità tecnologica che aspiri a essere autenticamente umana.

7. Conclusione

In conclusione, ciò che serve oggi non è un freno ideologico all'intelligenza artificiale, ma un investimento sistemico nella comprensione scientifica della tecnologia. Solo attraverso una piena padronanza dei modelli alla base dell'AI, una sperimentazione metodologica e una cultura tecnica condivisa, sarà possibile integrare questi strumenti nella società in modo efficace e sicuro. Non si tratta di sostituire l'intelligenza umana, ma di potenziarla, arricchendola con nuovi strumenti capaci di gestire la complessità del sapere contemporaneo. Come per ogni tecnologia dirompente, il futuro dell'AI non dipenderà da ciò che essa è oggi, ma da ciò che l'uomo sarà capace di farne, con conoscenza, responsabilità e spirito critico.

Bibliografia

Bishop, C. M. (1994). Neural networks and their applications. Review of Scientific Instruments, 65(6).

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6), December.

Chickering, D. M. (1996). Learning Bayesian Networks is NP-complete. In D. Fisher & H. J. Lenz (Eds.), Learning from Data: Artificial Intelligence and Statistics (Chap. 12). Springer Verlag.

Floridi, L. (2022). Etica dell'intelligenza artificiale: Sviluppi, opportunità, sfide. Raffaello Cortina Editore.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. Sage Journal, 61(4).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444.

MacKay, D. J. C. (2003). Information theory, inference, and learning algorithms (pp. 467-482). Cambridge University Press. (Pubblicazione del 2005)

Tosi, T. (2020). Qual è la differenza tra machine learning, deep learning e reti neurali? *SpremuteDigitali.com*. https://www.spremutedigitali.com.

APPRENDIMENTO DI ONTOLOGIE DEI CONTENUTI ABUSIVI ONLINE: SEEDED LDA E STRUTTURAZIONE SEMANTICA DELLE NARRAZIONI OFFENSIVE ANTI-MIGRANTI IN ITALIANO BASATA SU RETI

LEARNING ONTOLOGIES OF ONLINE ABUSIVE CONTENTS: SEEDED LDA AND GRAPH-BASED SEMANTIC STRUCTURING OF OFFENSIVE ANTI-MIGRANT NARRATIVES IN ITALIAN

Alex Cucco[1], Lara Fontanella[2], Annalina Sarra[3], Sara Fontanella[4]

Sommario

Le ontologie svolgono un ruolo fondamentale nella strutturazione e nell'interpretazione del vasto e non regolamentato discorso che caratterizza i dibattiti sociali online, in particolare su temi sensibili. Fornendo una comprensione formalizzata dei concetti e delle narrazioni che circolano negli ambienti digitali, le ontologie contribuiscono a rivelare schemi nascosti, chiarire significati ambigui e supportare analisi più approfondite del sentimento pubblico e delle cornici ideologiche. In questo studio proponiamo un approccio basato sui dati per supportare la costruzione di un'ontologia specifica di dominio, focalizzata sul dibattito online riguardante la migrazione nel contesto italiano. A partire da un corpus di 185,734 commenti generati dagli utenti estratti da Facebook, Instagram e YouTube, applichiamo tecniche esplorative per identificare schemi semanticamente significativi nel discorso pubblico. In particolare, utilizziamo un modello Seeded Latent Dirichlet Allocation per guidare il topic modeling mediante termini rilevanti per il dominio, integrato da un'analisi di rete semantica volta a mappare le relazioni tra i termini estratti. Questo approccio consente di individuare i concetti centrali e le loro interrelazioni all'interno delle discussioni online sulla migrazione. I risultati rappresentano la fase iniziale di un progetto più ampio volto allo sviluppo di risorse lessicali e ontologiche per migliorare l'identificazione automatica del linguaggio offensivo negli ambienti digitali e per caratterizzare le diverse narrazioni.

[1] University "G. d'Annunzio", Chieti-Pescara, Pescara, Italy - e-mail: alex.cucco@unich.it
[2] University "G. d'Annunzio", Chieti-Pescara, Pescara, Italy - e-mail: lara.fontanella@unich.it
[3] University "G. d'Annunzio", Chieti-Pescara, Pescara, Italy - e-mail: annalina.sarra@unich.it
[4] National Hearth and Lung Institute, Imperial College London, London, UK - e-mail: s.fontanella@imperial.ac.uk

Abstract

Ontologies play a critical role in structuring and interpreting the vast, unregulated discourse that characterizes online social debates, particularly on sensitive topics. By providing a formalized understanding of the concepts and narratives circulating in digital environments, ontologies help to uncover hidden patterns, clarify ambiguous meanings, and support deeper analyses of public sentiment and ideological framing. In this study, we propose a data-driven approach to support the construction of a domain-specific ontology focused on online debate on migration in the Italian context. Leveraging a corpus of 185,734 user-generated comments extracted from Facebook, Instagram, and YouTube, we apply exploratory techniques to identify semantically meaningful patterns in public discourse. Specifically, we use a Seeded Latent Dirichlet Allocation model to guide topic modeling with domain-relevant seed terms, complemented by semantic network analysis to map relationships among extracted terms. This approach enables the identification of core concepts and their interrelations within online discussions about migration. The results represent the initial phase of a broader project aimed at developing lexical and ontological resources to enhance automated detection of abusive languages in digital environments and to characterize different narratives.

Parole chiave: Narrative online, social media, ontologie, Seeded LDA, reti semantiche.

Keywords: *Online narratives, social media, ontology, Seeded LDA, semantic networks.*

1. Introduction

In the current digital era, online platforms serve as spaces for the circulation of information and the negotiation of public discourse on socially and politically sensitive issues. Within these environments, individuals express a broad spectrum of cultural, ideological, and personal viewpoints, often without constraints. The anonymity provided by social media platforms can encourage the unfiltered expression of opinions, including controversial or harmful viewpoints. This unaccountable environment fosters the circulation of emotionally charged, often offensive discourse (Tontodimamma *et al*., 2021, Fontanella *et al*., 2024a), including xenophobic and discriminatory narratives (Matamoros-Fernández and Farkas, 2021). New and old discriminatory practices are increasingly taking place on social media platforms. Importantly, much of this rhetoric does not always present as overt hate speech but often manifests subtly in what researchers refer to as ambient racism, a digital discursive form of discrimination that is pre-

sented as subtler, polished, and without obvious discriminatory language (Agudelo and Olbrych, 2022; Rubio-Carbonero, 2020). This phenomenon, as discussed by Sharma (2018), blurs the lines between orchestrated hate campaigns and the genuine emotional outbursts of ordinary users.

Identifying offensive online content is a fundamental application of event detection on social media. Although machine learning methods are frequently employed for automatic event detection, there has been an increasing interest in recent years in Explainable Artificial Intelligence (XAI) systems that offer comprehensible explanations for model decisions or predictions (for a recent review providing a structured taxonomy of XAI, see Schwalbe and Finzel 2024). Regardless of an AI model's accuracy or efficiency, users and practitioners frequently find it difficult to trust it if they cannot understand its functioning or the rationale behind its behavior. XAI aims to elucidate AI-generated prediction and decision using accessible language, so enhancing their transparency and reliability for human users (Adadi and Berrada, 2018). Incorporating human-centered explainability into event detection systems is essential for promoting more dependable and sustainable decision-making. For a decision to be comprehensible to humans, it must address the six essential inquiries – *who, what, when, where, why,* and *how* (often referred to as the 5W1H framework) (Kolajo and Daramola, 2023). Nonetheless, a significant difficulty exists in the incomplete nature of social media feeds, which frequently consist of unstructured, user-generated content lacking standardized structuring. The absence of structure complicates the extraction of all six aspects without dependence on external knowledge sources (Kolajo and Daramola, 2023). Consequently, to enhance the interpretability of XAI outputs and address informational deficiencies, it is necessary to integrate a framework that encompasses comprehensible concepts and explicitly delineated relationships among them. Ontologies, corresponding to conceptualization of a given domain, fulfill this role by offering structured definitions and interrelations that can substantiate the decisions made by XAI systems (Confalonieri *et al*., 2021, Confalonieri *et al*., 2024, Donadello and Dragoni 2021, Ribeiro and Leite 2021). The incorporation of contextual information from an ontology into an AI system promotes increased trust and trustworthiness in its resultant outcomes. Ontology learning is the process of deriving a structured representation of domain knowledge by extracting key concepts, their definitions, and the relationships among them from data. A notable subset of this field is Ontology Construction from Texts (OCT) (Tissaoui *et al*., 2022), which focuses specifically on extracting ontological elements from unstructured textual sources. This involves identifying terms, concepts, relationships, and axioms from text and using them to build or update ontologies (Wong *et al*., 2012). Traditionally, ontology construction comprises five core tasks: term extraction, concept formation, taxonomy derivation, identification of ad-hoc relationships, and axiom extraction (Asim

et al., 2018). Approaches to OCT can be broadly categorized into statistical, linguistic, logic-based, or hybrid techniques (Wong *et al*., 2012). Statistical methods, operating primarily at the syntactic level, are especially useful in the early stages of ontology learning, such as during term identification and hierarchy formation. These methods rely on the premise that word co-occurrence patterns offer meaningful insights into semantic relationships. Among the statistical techniques, Latent Dirichlet Allocation (LDA) has been employed in ontology construction (Colace *et al*., 2016, Tissaoui *et al*., 2020, Zhang *et al*., 2019). For instance, Colace *et al*. (2016) utilized LDA to extract domain-relevant terms and their co-occurrence patterns from a set of domain-specific documents, producing a graph that represents term relationships. Building on the strengths of clustering methods in concept formation, Huang *et al*. (2021) proposed a seed-knowledge-based LDA approach, where topics are initialized with predefined seed terms linked to core concepts, thereby guiding the model to form semantically coherent clusters aligned with domain knowledge.

In our study, we focus on constructing an ontology from textual corpora, with particular attention to extracting relevant terms associated with the domain of online debate about migration and identifying the relationships among abusive extracted keywords. Our analysis adopts a statistical approach, specifically employing a Seeded LDA model (Jagarlamudi *et al*., 2012), to guide the extraction and clustering of semantically related terms, and a semantic network to study the relations between the relevant terms. This procedure is applied to a corpus of user-generated comments concerning migration and migrants, collected from social media platforms including Facebook, Instagram, and YouTube. This work represents the initial phase of a broader project that aims to develop comprehensive lexical resources and ontological frameworks to support the automated detection of online abusive language and xenophobia and extend the results presented in Fontanella *et al*. (2024b).

2. Data collection and preprocessing

To extract the keywords related to public sentiment surrounding migration in the Italian context, we assembled a comprehensive corpus of user-generated content from three major social media platforms: Facebook, Instagram, and YouTube. These platforms were selected due to their wide user base, heterogeneous demographic representation, and the interactive nature of their comment sections, which offer valuable insight into spontaneous, real-time public opinion. For each platform, posts or videos addressing immigration-related topics were manually selected, and the corresponding user comments were retrieved using the exportcomments.com tool. The resulting corpus comprises 185,734 comments – 129,602 from Facebook, 4,749 from Instagram, and 51,383 from YouTube – shared on these platforms over the past ten years. Before

initiating the analytical phase, we subjected the dataset to a thorough preprocessing pipeline to enhance the consistency and relevance of the textual material. Standard natural language processing (NLP) techniques were applied, including tokenization, normalization, removal of stopwords, and filtering of irrelevant or noisy data. These steps were essential to minimize linguistic redundancy and maximize semantic clarity.

The refined dataset was processed using the R package *quanteda* (Benoit *et al*., 2018), which facilitated efficient text manipulation and feature extraction. As a result, the final corpus was distilled into a vocabulary of 9,060 unique tokens.

3. Semi-Supervised Topic Modeling: Seeded LDA

Topic modeling remains one of the most effective methods for clustering words on large textual corpora according to latent themes. Among these, LDA (Blei and Jordan 2003) is widely used, modeling each document as a mixture of topics, and each topic as a distribution over words.

However, our primary focus is on term clustering approaches for concept generation, with the ultimate objective of developing an ontology specific to the domain of online abusive contents related to migration and migrants. Specifically, we aim to group terms according to the core concepts of a domain ontology to guarantee semantic consistency and domain pertinence (Huang *et al*., 2021). To this end, we adopt the Seeded LDA framework (Jagarlamudi *et al*., 2012), which extends conventional topic modeling by incorporating prior knowledge through the use of seed terms – words explicitly associated with predefined domain concepts. The key innovation of Seeded LDA lies in its ability to guide the topic formation process toward semantically meaningful clusters. This is achieved by introducing a set of seeded topics that are initialized with predefined terms and by modifying the generative process to alternate between regular, data-driven topic assignments and seed-guided topic assignments. Specifically, each word token in a document is assigned to one of two topic distributions, a regular topic distribution and a seed-based topic distribution, with the selection controlled by a Bernoulli random variable. The model also incorporates an additional Dirichlet prior informed by binary vectors indicating the presence of seed words, thus constraining topic-word probabilities around the predefined conceptual anchors. For a comprehensive description of the technical aspects, see Jagarlamudi *et al*. (2012).

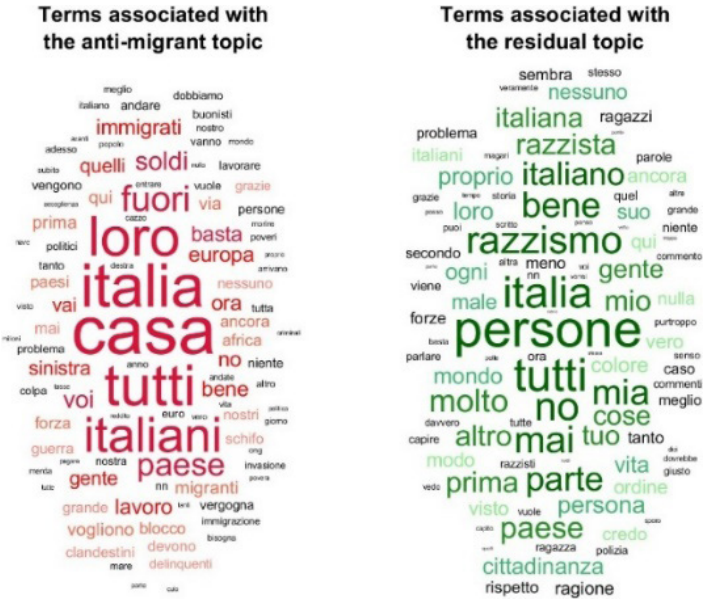
In our application, we implemented Seeded LDA by defining one seed-guided topic driven by a curated set of seed terms extracted from the "negative stereotype ethnic slurs" category of the revised HurtLex lexicon (Tontodimamma *et al*., 2023), alongside an additional unguided (residual) topic to capture general discourse not directly related to the predefined conceptual space. This configuration allows the model to steer the allocation of semantically related words toward the domain-relevant topic while preserving

flexibility to discover emergent linguistic patterns in the remaining data. The resulting topic-word distribution for the seed-guided topic forms the basis for subsequent concept identification and ontology construction. In fact, this setup enables the discovery of new terms that are contextually associated with the seed words, thereby expanding the original lexicon in a data-driven yet controlled manner.

3.1 The extended domain specific lexicon from Seeded LDA

Given the corpus of comments and the selected seed words for the anti-migrant topic, we implemented the Seeded LDA model using the quanteda R package (Benoit *et al*., 2018). In this approach, each document is modeled as a mixture of a seed-guided topic, biased toward the predefined set of terms associated with offensive anti-migrant discourse, and a residual topic capturing general or unrelated content. To classify comments as pertaining to the offensive anti-migrant topic, a probabilistic thresholding approach was employed: a document was assigned to the seed-guided topic if its posterior probability for that topic exceeded 0.5, otherwise it was assigned to the residual topic. Based on this criterion, approximately 44.4% of the comments were classified under the seed-guided topic. Figure 1 illustrates the lexical composition of both topics, displaying the top 100 terms most strongly associated with each.

Figure 1. Word clouds displaying the top 100 terms associated with the anti-migrant topic and the residual topic. Word size reflects the word-topic probabilities estimated using the Seeded LDA model



In the offensive anti-migrant cluster, terms such as *loro* (they), *casa* (home), and *Italia*/*Italiani* (Italy/Italians) dominate, reflecting a "us versus them" rhetoric combined with national identity and exclusionary sentiment. Conversely, the residual cluster is characterized by terms like *persone*/*persona* (persons/person), *razzismo*/*razzista* (racism/racist) generally linked to a more empathetic or reflective tone, often centered around universalist or anti-racist discourse.

To further investigate the textual content associated with the anti-migrant topic, we categorized the corresponding comments into five classes based on their document-topic probability scores relative to the seed-guided topic. Table 1 presents the distribution of comments across these classes, both for the entire corpus and disaggregated by platform. Notably, comments exhibiting a high degree of alignment with the offensive anti-migrant topic (probability > 0.8) constitute 25% of the dataset, although the distribution varies across platforms. These findings suggest platform-specific dynamics in the expression of offensive anti-migrant discourse, with indications of thematic polarization or content divergence emerging in relation to the guided topic.

Table 1. Distribution of comments across probability classes for assignment to the seed-guided anti-migrant topic, disaggregated by platform

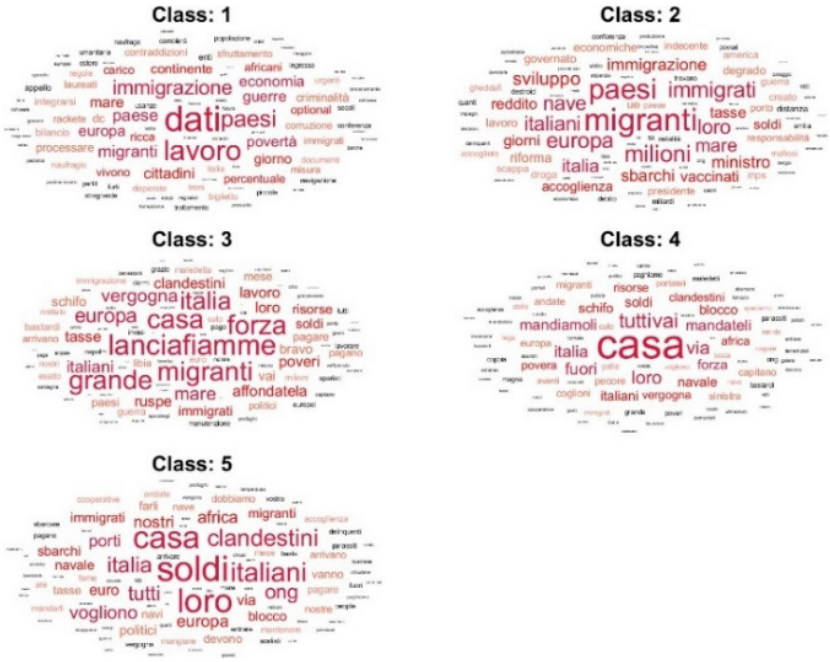
	Class	Probability thresholds	Corpus	Facebook	Instagram	YouTube
Residual topic		*<0.5*	*55.6 %*	*52.0 %*	*35.6 %*	*66.5 %*
Seed-guided anti-migrant topic	1	0.5-0.6	3.0%	3.0%	4.0%	3.0%
	2	0.6-0.7	6.5%	6.8%	9.7%	5.5%
	3	0.7-0.8	9.9%	10.6%	11.5%	8.2%
	4	0.8-0.9	15.7%	17.6%	21.9%	10.4%
	5	>0.9	9.3%	10.1%	17.2%	6.5%

Figure 2 disaggregates the anti-migrant narrative by visualizing the keywords associated with the sub-corpora of documents corresponding to the five classes. The keywords were identified through keyness analysis (Gabrielatos, 2018), a statistical method designed to identify terms that appear significantly more often in one sub-corpus compared to the other sub-corpora. A keyness score was determined through the chi-square test. This score evaluates whether the actual frequency of a word in the target corpus significantly deviates from the expected frequency derived from the reference corpus. We used a statistical significance threshold (*p*-value for the chi-square test below 0.001) to identify significantly over-represented words in each subgroup.

As the probability that the documents are assigned to the offensive anti-migrant topic increases from 1 to 5, a clear intensification in both lexical focus and emotional charge is observed. At the lower level (Class 1), the terms reflect a more neutral or factual discourse centered on structural and geopolitical themes, such as migration, poverty, economy, and war, rather than overt hostility. The vocabulary suggests a broader discussion on migration as a socioeconomic phenomenon rather than a polarizing or accusatory narrative. For the documents assigned to Class 2, the terms blend references to migration (*migranti/ immigrati* – migrants/ immigrants, *sbarchi* – landings, *accoglienza* – reception system) with politically and economically charged terms (*tasse* – taxes, *reddito* – income, *soldi* –money, *governato* – governed, *riforma* – reform, *ministro* – minister, *UE* – EU). This suggests a discourse that, while not overtly hostile, reflects concerns about national governance, economic impact, and public resources, often framing migration within debates on state management and social burden. The terms also point to criticism or questioning of political handling of migration. At higher assignment levels (Classes 3 through 5), terms like *mandiamoli/mandateli fuori* (send them away), *clandestini* (illegals), and *lanciafiamme* (flamethrower) *parassiti* (parasites) become more prominent. These words suggest a shift toward increasingly hostile and exclusionary discourse, emphasizing national sovereignty, resource competition, and border control. The presence of adversarial terms such as *schifo* (disgust), *blocco* (blockade), and *porti* (ports) at the highest level (Classe 5) further underscores the escalating tone of rejection and perceived threat.

Together, Figures 1 and 2 highlight the lexical distinctions between xenophobic and neutral narratives, as well as the continuum of hostility reflected in increasing document-topic probabilities. This progression illustrates how Seeded LDA facilitates the identification and clustering of domain-specific concepts related to offensive anti-migrant discourse, thereby supporting the systematic construction of an ontology that captures varying degrees of discriminatory language. By operationalizing both explicit and implicit forms of online discrimination, this approach helps improve the semantic richness and granularity of the resulting ontology within the domain of online abusive content.

Figure 2. Word clouds illustrating the keywords within subgroups of comments, segmented by different threshold levels of document-topic probabilities for the offensive anti-migrant topic

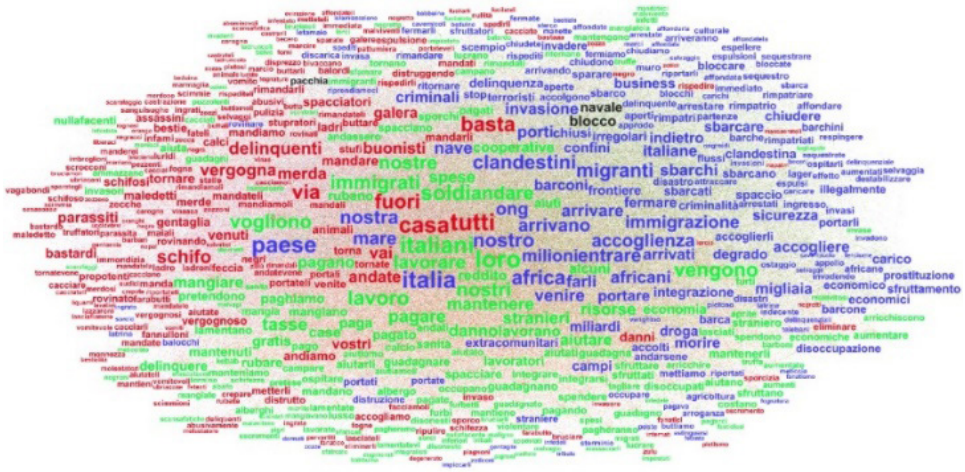


4. Semantic network

Focusing on the keywords characterizing the five sub-corpora associated with abusive anti-migrant narratives, we constructed a semantic network to explore the relationships among salient terms. In this network, nodes represent keywords, and edges capture the degree of semantic relatedness between them. The resulting graph structure enables the representation and analysis of associations and dependencies among key concepts within the domain. As discussed in Pronello *et al*. (2024), several approaches can be used to define links in a semantic graph; among these, co-occurrence-based methods are widely adopted and form the basis of our analysis. This corpus-based measure of semantic relatedness relies on the distributional hypothesis, which posits that words appearing in similar contexts tend to share similar meanings (Harispe *et al*., 2015). To build the semantic network, we computed pairwise co-occurrence statistics among keywords and used these values to establish weighted links in the graph. The final network consists of 624 nodes with a density of 0.366. To identify semantically coherent clusters within the network, we employed the Louvain community detection algorithm (Blondel *et al*., 2008), which efficiently partitions large graphs by maximizing modularity. Through an iterative, hierarchical process, the algorithm groups densely connected nodes, enabling

the identification of semantically related word clusters. Figure 3 displays the resulting semantic graph, where node colors indicate community membership as identified by the Louvain algorithm, and word size is scaled according to term frequency. The resulting communities highlight meaningful groupings of co-occurring terms, revealing the internal organization and thematic diversity of hostile discourse toward migrants.

Figure 3. Semantic Network of the characteristic terms of the anti-migrant narratives identified through keyness analysis



One prominent cluster (in red) represents the hostile core of anti-migrant discourse, comprising a cohesive set of highly charged abusive terms identified within the semantic network. These terms form the foundational vocabulary of hostility, exclusion, and dehumanization in online xenophobic and racist rhetoric. Characterized by strong emotive and aggressive content, this lexicon includes ethnic slurs, dehumanizing metaphors, expulsion imperatives, and explicit incitements to violence (see Table 2). The second cluster (e.g., in green) revolves around socio-economic tension and perceived resource competition, reflecting a narrative of economic resentment toward migrants. This lexical community is built around verbs and nouns related to work, welfare, public spending, and entitlement, and conveys the belief that migrants unfairly benefit from resources funded by native citizens. This cluster sustains anti-migration narratives in which migrants are viewed as exploiters, burdens, or undeserving beneficiaries, and contributes to the moral and economic justification for exclusionary attitudes. Finally, a third distinct cluster (in blue) captures the geopolitical and logistical framing of migration, emphasizing the scale, modes, and perceived consequences of migrant arrivals. This cluster includes references to territories, migration flows, and modes of entry. It

also contains terms related to actors and reception infrastructure, as well as expressions of border control and security concerns. In the ontology, this cluster informs categories related to border and territorial integrity, population pressure, institutional handling of migration, and security discourse, often aligned with alarmist or defensive narratives about migration.

The semantic segmentation enabled by this analysis reveals the multidimensional architecture of anti-migrant rhetoric and offensive anti-migrant rhetoric, which ontology construction helps to formalize and interpret. Rather than being monolithic, such rhetoric is composed of intersecting sub-narratives involving identity, economy, criminality, and geopolitics. By applying a network-based clustering approach, we can identify coherent communities of meaning and classify terms according to their discursive function. This method supports the development of a structured ontology that captures how different forms of anti-migrant discourse and offensive anti-migrant discourse are lexically constructed and discursively sustained in online environments.

Table 2. Categories of abusive terms identified in offensive anti-migrant discourse, with examples illustrating their semantic functions and typical expressions in Italian and English

Term categories	Examples
Violent Threats – explicitly incite harm or elimination.	*sparategli* (shoot them), *bruciamoli* (burn them), *fucilateli* (execute them by firing squad).
Dirt and Filth Metaphors – describe migrants as waste or pollution, emphasizing disgust.	*immondizia* (garbage), *luridi* (filthy), *vomitevoli* (vomit-inducing).
Animalistic and Dehumanizing Labels – strip migrants of human qualities.	*bestie* (beasts), *scimmie* (monkeys), *scarafaggio* (cockroach).
Criminality and Moral Corruption – includes accusations of deviance or threat to order.	*ladri* (thieves), *malviventi* (criminals), *molestatori* (molesters).
Expulsion and Removal Imperatives – urge action to eject or eliminate migrants.	*cacciateli* (kick them out), *rimandateli* (send them back), *buttiamoli* (let's throw them out).
Parasites and Exploiters – accuse migrants of draining resources or being unproductive.	*parassiti* (parasites), *sanguisughe* (leeches), *fannulloni* (idlers).
Extremely offensive Racialized Slurs	*negri (*niggers*), negretti (*little niggers*), negracci (*dirty niggers*), zulù (*Zulu used pejoratively*).*
Calls for Collective Action or Genocide: incite group violence or extermination with military or genocidal imagery	*forni (*ovens*), lanciafiamme (*flamethrower*), uccidetevi (*go kill yourselves*), massacrateli (*massacre them*).*
Nationalistic Rhetoric: frame migrants as intruders who contaminate national identity	*invasori (*invaders*), invaso (*invaded*), contaminati (*contaminated*).*
Insults and General Contempt – broad derogatory terms expressing disdain.	*bastardi* (bastards), *feccia* (scum), *gentaglia* (rabble).

5. Conclusion and future work

This study demonstrates how semi-supervised topic modeling and semantic network analysis can uncover the rhetorical structures of anti-migrant discourse and abusive anti-migrant discourse in Italian social media. By distinguishing clusters of offensives versus neutral narratives, we validate the effectiveness of Seeded LDA in capturing domain-specific patterns and linguistic markers of online xenophobia. The semantic segmentation of the discourse reveals that anti-migrant rhetoric is not monolithic but organized into overlapping thematic dimensions, including different levels of offensiveness that were described in this work, from racialized dehumanization to economic resentment, criminalization, and nationalistic fear, each characterized by distinct lexical fields. The integration of co-occurrence-based graph clustering further enhances our understanding of how key terms operate as semantic anchors within hostile narratives. This dual approach offers a scalable and data-driven foundation for ontology learning. This initial analysis of textual data represents a foundational step toward the longer-term objective of constructing a structured ontology of abusive and xenophobic speech. However, a limitation of this initial analysis is the absence of a semantic disambiguation step, which we plan to address by integrating ontology-aware word-in-context disambiguation (Martelli *et al*., 2021) and taxonomic reasoning (Camacho-Collados *et al*., 2018). In addition, given that an ontology serves as a formal representation of concepts, their interrelations, and attributes within a domain, we aim to structure the extracted vocabulary and semantic relationships into a formal ontology using established standards such as SKOS (Simple Knowledge Organization System) and OWL (Web Ontology Language). SKOS supports the representation of lightweight knowledge organization systems, enabling concepts to be linked by hierarchical (broader/narrower) and associative (related) relationships. This is useful for organizing anti-migrant terms into coherent categories and showing their semantic proximity. OWL provides greater expressiveness, allowing for the formalization of complex class hierarchies, property restrictions, and logical axioms. This enables the ontology to represent, for example, that "racial slurs" are a subclass of "abusive terms", or that certain terms are exclusively associated with specific ethnic groups. By leveraging these semantic web standards, the ontology will be interoperable, extensible, and machine-executable, facilitating integration into computational frameworks for abusive content detection, content moderation, and explainable AI. Specifically, the ontology knowledge could be incorporated at various stages of the machine learning pipeline, including ontology-based feature engineering, algorithm design, training, and interpretation, thus enhancing transparency and accountability in automated racism detection (Ghidalia *et al*., 2024).

Acknowledgments. This work is part of the research project PRIN-2022 PNRR Identification and Critical Analysis of Online Racism and Xenophobia against (Im)migrants and Roma people (Project Code: P2022APKJL), funded by the European Union – Next Generation EU.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agudelo, F. I., & Olbrych, N. (2022). It's not how you say it, it's what you say: Ambient digital racism and racial narratives on Twitter. *Social Media + Society, 8*(3). https://doi.org/10.1177/20563051221122441

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. *Database, 2018*, bay101. https://doi.org/10.1093/database/bay101

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774. https://doi.org/10.21105/joss.00774

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Camacho-Collados, J., Delli Bovi, C., Espinosa-Anke, L., Oramas, S., Pasini, T., Santus, E., Shwartz V., Navigli R., & Saggion. H. (2018). SemEval-2018 Task 9: Hypernym Discovery. Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), 712-724.

Colace, F., De Santo, M., Greco, L., Moscato, V., & Picariello, A. (2016). Probabilistic approaches for sentiment analysis: Latent Dirichlet allocation for ontology building and sentiment extraction. In W. Pedrycz & S.-M. Chen (Eds.), *Sentiment analysis and ontology engineering Studies in Computational Intelligence* (Vol. 639, pp. 57-76). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-30319-2_4

Confalonieri, R., Kutz, O., Calvanese, D., Alonso-Moral, J. M., & Zhou, S.-M. (2024). The role of ontologies and knowledge in explainable AI: Editorial. *Semantic Web, 15*(4), 933-936. https://doi.org/10.3233/SW-243529

Confalonieri, R., Weyde, T., Besold, T. R., & Martin, F. M. (2021). Using ontologies to enhan-

ce the understandability of global post-hoc explanations of black-box models. *Artificial Intelligence, 296*, 103471. https://doi.org/10.1016/j.artint.2021.103471

Donadello, I., & Dragoni, M. (2021). SeXAI: A semantic explainable artificial intelligence framework. In M. Baldoni & S. Bandini (Eds.), *AIxIA 2020. Advances in artificial intelligence* (Vol. 12414, pp. 41-55). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-77091-4_4

Fontanella, L., Chulvi, B., Ignazzi, E., Sarra, A., & Tontodimamma, A. (2024a). How do we study misogyny in the digital age? A systematic literature review using a computational linguistic approach. *Humanities and Social Sciences Communications, 11*, 478. https://doi.org/10.1057/s41599-024-02978-7

Fontanella, L., Sarra, A., del Gobbo, E., Cucco, A., & Fontanella, S. (2024b). Exploring anti-migrant rhetoric on Italian social media. In A. Plaia, L. Egidi, & A. Abbruzzo (Eds.), *Proceedings of the SDS 2024 Conference: New perspectives on statistics and data science* (pp. 45-60). Palermo, Italy: Università degli Studi di Palermo.

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse* (pp. 225-258). London, England: Routledge. https://doi.org/10.4324/9781315179346-11

Ghidalia, S., Labbani Narsis, O., Bertaux, A., & Nicolle, C. (2024). Combining machine learning and ontology: A systematic literature review. *arXiv preprint* arXiv:2401.07744. https://arxiv.org/abs/2401.07744

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). *Semantic similarity from natural language and ontology analysis*. San Rafael, CA: Morgan & Claypool. https://doi.org/10.2200/S00639ED1V01Y201504HLT027

Huang, H., Harzallah, M., Guillet, F., & Xu, Z. (2021). Core-concept-seeded LDA for ontology learning. *Procedia Computer Science, 192*, 222-231. https://doi.org/10.1016/j.procs.2021.08.023

Jagarlamudi, J., Daumé, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204-213). Stroudsburg, PA: Association for Computational Linguistics.

Kolajo, T., & Daramola, O. (2023). Human-centric and semantics-based explainable event detection: A survey. *Artificial Intelligence Review, 56*(Suppl. 1), 119-158. https://doi.org/10.1007/s10462-023-10525-0

Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021). SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). *Proceedings of the 15th International Workshop on Semantic Evaluation*, 24-36.

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.

org/10.1177/1527476420982230

Pronello, N., Cucco, A., del Gobbo, E., & Fontanella, L. (2024). Dynamics of online debates: Insights from textual network analysis. *Annals of Operations Research*. https://doi.org/10.1007/s10479-024-06315-8

Ribeiro, M. S., & Leite, J. (2021). Aligning artificial neural networks and ontologies towards explainable AI. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(6), 4932-4940. https://doi.org/10.1609/aaai.v35i6.16626

Rubio-Carbonero, G. (2020). Subtle discriminatory political discourse on immigration. *Journal of Language and Politics, 19*(6), 894-915. https://doi.org/10.1075/jlp.19069.rub

Schwalbe, G., & Finzel, B. (2024) A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38, 3043-3101 (2024). https://doi.org/10.1007/s10618-022-00867-8

Sharma, C. (2018). Affect and the attention economy of online racism. In *Proceedings of the 67th Annual Conference of the International Communication Association* (pp. 1-10). Prague, Czech Republic: International Communication Association.

Tissaoui, A., Sassi, S., & Chbeir, R. (2020). Probabilistic topic models for enriching ontology from texts. *SN Computer Science, 1*, 336. https://doi.org/10.1007/s42979-020-00349-y

Tissaoui, A., Sassi, S., Chbeir, R., & Mechergui, A. (2022). A top-down enriching approach for ontology learning from text. *Concurrency and Computation: Practice and Experience, 34*(19), e7036. https://doi.org/10.1002/cpe.7036

Tontodimamma, A., Fontanella, L., Anzani, S., & Basile, V. (2023). An Italian lexical resource for incivility detection in online discourses. *Quality & Quantity, 57*, 3019-3037. https://doi.org/10.1007/s11135-022-01494-7

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics, 126*, 157-179. https://doi.org/10.1007/s11192-020-03737-6

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys, 44*(4), 20:1-20:36. https://doi.org/10.1145/2333112.2333115

Zhang, Q., Liu, S., Gong, D., & Tu, Q. (2019). A latent-Dirichlet-allocation-based extension for domain ontology of enterprise's technological innovation. *International Journal of Computers Communications & Control, 14*, 107-123. https://doi.org/10.15837/ijccc.2019.1.3366

SINTETIZZARE LA CONOSCENZA: UN APPROCCIO INTEGRATO PER L'ESTRAZIONE DI CONTENUTI RILEVANTI NELLA LETTERATURA SCIENTIFICA

SYNTHESIZING KNOWLEDGE: AN INTEGRATED APPROACH FOR EXTRACTING RELEVANT CONTENT FROM SCIENTIFIC LITERATURE

Massimo Aria[1,4], Corrado Cuccurullo[2,4], Luca D'Aniello[1,4], Michelangelo Misuraca[3,4], Maria Spano[1,4]

Sommario

L'aumento esponenziale della produzione scientifica rende sempre più complessa l'individuazione rapida dei contributi rilevanti nella letteratura. In questo contesto, i metodi di sintesi automatica dei testi offrono soluzioni promettenti, permettendo la generazione di riassunti informativi da documenti lunghi e strutturati. Questo studio introduce l'*Integrated Text Summarization* (ITS), un nuovo approccio estrattivo non supervisionato progettato specificamente per i testi scientifici. L'algoritmo combina l'analisi strutturale del documento con l'integrazione di parole chiave fornite dagli autori e/o estratte automaticamente dal testo, al fine di selezionare le frasi più rilevanti in ciascuna sezione. L'ITS è stato valutato su un campione multidisciplinare di articoli, confrontando i risultati con frasi indicate dai loro autori. Le prestazioni sono state inoltre messe a confronto con due metodi di riferimento: l'algoritmo TextRank e il modello GPT-4o. I risultati mostrano che l'ITS raggiunge una maggiore accuratezza e stabilità nella selezione dei contenuti rilevanti, anche in contesti disciplinari diversi. L'approccio si configura quindi come una soluzione trasparente, interpretabile ed efficace per la sintesi automatica della conoscenza scientifica.

Abstract

The exponential growth of scientific production has made it increasingly difficult to rapidly identify the most relevant contributions within the literature. In this con-

[1] Università di Napoli "Federico II", Dipartimento di Scienze Economiche e Statistiche, Napoli, Italia - e-mail: massimo.aria@unina.it; luca.daniello@unina.it (corresponding author); maria.spano@unina.it.

[2] Università degli Studi della Campania "Luigi Vanvitelli", Dipartimento di Economia e Management, Capua, Italia - e-mail: corrado.cuccurullo@unicampania.it.

[3] Università degli Studi di Salerno, Dipartimento di Scienze Aziendali - Management & Innovation Systems, Fisciano, Italia - e-mail: mmisuraca@unisa.it.

[4] Università di Napoli "Federico II", K-Synth Spin-Off, Napoli, Italia.

text, automatic text summarization methods offer promising solutions, enabling the generation of informative summaries from long and structured documents. This study introduces Integrated Text Summarization *(ITS), a novel unsupervised extractive approach specifically designed for scientific texts. The algorithm combines structural analysis of the document with the integration of keywords provided by the authors and terms automatically extracted from the text, in order to identify the most relevant sentences in each section. ITS was evaluated on a multidisciplinary sample of scientific articles by comparing the extracted sentences with those selected by the original authors. Its performance was further benchmarked against two reference methods: the classical TextRank algorithm and the generative model GPT-4o. The results show that ITS achieves greater accuracy and stability in identifying relevant content, even across diverse disciplinary domains. The proposed approach thus emerges as a transparent, interpretable, and effective solution for the automatic summarization of scientific knowledge.*

Parole chiave: Sintesi automatica del testo; Estrazione di informazioni; keyword scientifiche; LLMs; metodi estrattivi non supervisionati.
Keywords: *Automatic Text Summarization; Information extraction; Scientific keywords; LLMs; Extractive summarization.*

1. Introduzione

Il sovraccarico informativo, caratterizzato da una crescita esponenziale dei dati testuali, coinvolge un numero crescente di domini disciplinari (Sarker *et al*., 2017), con particolare incidenza nell'ambito della letteratura scientifica (Landhuis, 2016). La rapida proliferazione di articoli ha reso sempre più difficile per i ricercatori identificare i contributi più rilevanti e coglierne rapidamente le implicazioni teoriche e pratiche. In questo contesto, si avverte pertanto la necessità di sviluppare strumenti in grado di aumentare l'efficienza dei processi di recupero e sintesi dell'informazione, riducendo il carico cognitivo e il tempo richiesto per la consultazione.

Tra le tecniche più promettenti si colloca la Sintesi Automatica dei Testi (Automatic Text Summarization, ATS), che consente la generazione di riassunti compatti ed esaustivi a partire da documenti complessi, facilitando l'accesso rapido alle informazioni più rilevanti. I riassunti prodotti includono i concetti chiave del testo originale, minimizzando la ridondanza e preservando l'integrità semantica del documento.

A partire da queste premesse, si propone un nuovo metodo di sintesi automatica progettato specificamente per articoli scientifici. L'approccio sviluppato integra caratteristiche peculiari delle pubblicazioni accademiche per individuare e selezionare frasi chiave lungo l'intero documento. Testato su articoli appartenenti a discipline differenti,

il metodo ha dimostrato una capacità superiore di identificazione delle frasi significative rispetto sia a TextRank, uno dei metodi di sintesi automatica più utilizzati, sia ai modelli di Large Language Model (LLM), come GPT-4o integrato in ChatGPT, un avanzato modello generativo, contribuendo a una sintesi più accurata e informativa dei contenuti.

2. Sintesi automatica dei testi: stato dell'arte, vantaggi e limiti nei testi scientifici

Le tecniche di ATS si suddividono principalmente in due approcci: estrattivo e astrattivo. L'estrattivo seleziona frasi direttamente dal testo originale in base alla loro rilevanza, copiandole senza modifiche nel riassunto finale. Quello astrattivo, invece, riformula i contenuti creando nuove frasi che condensano le idee principali (Nenkova e McKeown, 2011).

L'approccio estrattivo presenta alcuni vantaggi distintivi rispetto a quello astrattivo. In particolare:

1. garantisce una maggiore accuratezza fattuale, poiché utilizza direttamente le frasi originali, riducendo il rischio di errori e distorsioni;
2. è più efficiente dal punto di vista computazionale, risultando idoneo per applicazioni in tempo reale e per l'elaborazione di grandi volumi di dati (Gambhir e Gupta, 2017);
3. richiede meno dati di addestramento ed è più semplice da implementare, favorendone la diffusione in ambiti applicativi e di ricerca eterogenei;
4. conserva meglio lo stile e l'intento dell'autore, caratteristica particolarmente rilevante nei contesti scientifici e tecnici, dove precisione e fedeltà espressiva risultano essenziali (Mani, 2001).

L'applicazione della sintesi automatica alla letteratura scientifica è stata ampiamente esplorata (Zaheer *et al*., 2020; Koh *et al*., 2022). Tuttavia, sebbene molti studi si siano concentrati sulla riduzione della lunghezza dei testi, la sintesi di articoli scientifici presenta ancora numerose criticità. Anche utilizzando modelli linguistici più avanzati, come BART (Lewis *et al*., 2019) e PEGASUS (Zhang *et al*., 2020), l'elaborazione dei contenuti degli articoli risulta essere complessa e limitata.

I recenti sviluppi nel campo del deep learning, in particolare l'introduzione dei modelli *transformer* come BERT e GPT, hanno migliorato sensibilmente le prestazioni nella sintesi testuale, mostrando notevoli capacità di comprensione e generazione del linguaggio naturale (Brown *et al*., 2020). Tuttavia, la sintesi di documenti scientifici lunghi comporta sfide aggiuntive, quali l'identificazione di contenuti chiave dispersi tra le varie sezioni del testo. Per essere davvero efficaci, i metodi di sintesi devono dunque tener conto della struttura globale del documento. Alcuni approcci recenti (Chen e Bansal, 2018; Meng *et al*., 2021) affrontano questo aspetto, ma rimangono fortemente one-

rosi in termini computazionali e spesso faticano a estrarre con precisione la conoscenza essenziale, proprio a causa della natura strutturata dei testi scientifici.

3. Un approccio integrato di sintesi automatica dei testi scientifici

Alla luce delle attuali limitazioni dei metodi di sintesi automatica astrattiva, in particolare per l'elevata complessità computazionale e il rischio di generare contenuti semanticamente imprecisi, il presente studio propone un metodo basato su tecniche estrattive. Tali metodi possono essere classificati in tre principali categorie: (1) approcci statistici, basati su misure quantitative come la frequenza di termini o la similarità lessicale; (2) approcci basati su regole, che utilizzano criteri predefiniti per la selezione delle frasi; e (3) approcci fuzzy, che impiegano principi di logica fuzzy per valutare la rilevanza informativa.

Nell'ambito degli approcci statistici, rivestono un ruolo centrale i modelli basati su network di similarità. In tali configurazioni, ogni frase del testo è formalizzata come un nodo all'interno di un grafo; gli archi connettono i nodi qualora venga superata una specifica soglia di similarità, calcolata mediante sovrapposizione lessicale o prossimità semantica (valutata, ad esempio, attraverso i *word embedding*). In questa struttura grafica, la centralità dei nodi costituisce un indicatore chiave: le frasi con i punteggi di centralità più elevati vengono selezionate per il riassunto.

Due degli algoritmi estrattivi basati su network di similarità più consolidati nella letteratura sono TextRank (Mihalcea & Tarau, 2004) e LexRank (Erkan & Radev, 2004).

TextRank calcola la similarità tra frasi utilizzando l'*overlap* normalizzato di parole, pesato rispetto alla lunghezza delle frasi, e costruisce un grafo in cui i nodi sono connessi se la similarità supera una certa soglia. Su questa rete viene poi applicato l'algoritmo PageRank (un metodo matematico che assegna un punteggio di importanza a ciascun nodo della rete in base al numero e alla qualità delle connessioni che riceve), originariamente sviluppato per classificare l'importanza delle pagine web nel motore di ricerca Google. Il punteggio di centralità calcolato da PageRank riflette l'importanza relativa di ciascuna frase nel contesto del documento, permettendo di selezionare quelle più informative per la sintesi.

Pur mantenendo un'impostazione analoga al TextRank, LexRank si distingue per l'utilizzo della similarità del coseno pesata su vettori TF-IDF. Tale schema valuta non solo la frequenza dei termini, ma anche la loro capacità discriminativa all'interno del corpus. Sebbene questa metrica garantisca una rappresentazione semantica più raffinata, essa comporta un onere computazionale notevole, necessitando il calcolo della similarità per ogni possibile coppia di frasi (calcolo *pairwise*). Entrambi gli algoritmi si basano su misure di centralità topologica per ordinare le frasi secondo la loro rilevanza. Tuttavia, considerati i vincoli computazionali legati all'elaborazione di documenti scientifici

di grandi dimensioni, l'approccio di sintesi proposto in questo lavoro adotta il TextRank come nucleo metodologico. Esso rappresenta un compromesso efficace tra prestazioni computazionali e capacità di identificare frasi salienti, risultando particolarmente adatto a scenari di applicazione su larga scala.

Il metodo di sintesi automatica dei testi proposto, denominato *Integrated Text Summarization* (ITS), si distingue dai metodi estrattivi esistenti per essere un approccio non supervisionato progettato *ad hoc* per la specificità dei documenti scientifici. I documenti scientifici si caratterizzano per una struttura testuale formalizzata e fortemente organizzata, che segue nella maggior parte dei casi lo schema IMRaD (*Introduction, Methods, Results, and Discussion*), ovvero l'organizzazione standard degli articoli empirici in cui l'introduzione presenta il problema di ricerca, i metodi descrivono come è stato condotto lo studio, i risultati riportano i dati raccolti, e la discussione interpreta i risultati alla luce della letteratura esistente. Questa articolazione in sezioni distinte riflette una sequenza logico-argomentativa che facilita non solo la lettura e la comprensione da parte dei lettori, ma anche l'analisi computazionale del contenuto. Proprio a partire da questa osservazione si sviluppa la prima caratteristica chiave dell'approccio ITS: l'applicazione dell'algoritmo di sintesi in modalità sezionale, ovvero mediante l'analisi autonoma di ciascuna sezione del documento. Attraverso questa strategia, ITS è in grado di identificare frasi chiave lungo tutto l'arco del testo, garantendo una copertura informativa bilanciata e coerente rispetto alla struttura originaria dell'articolo. Inoltre, lo sviluppo dell'ITS basato su algoritmo non supervisionato, ne rafforza la versatilità e l'applicabilità in contesti multidisciplinari, senza necessità di addestramento preventivo su corpora annotati.

Un aspetto particolarmente innovativo dell'ITS riguarda l'integrazione delle parole chiave nel processo di sintesi, in quanto considerate indicatori privilegiati dei concetti fondamentali espressi nel documento. Le parole chiave vengono generalmente selezionate dagli autori con estrema attenzione, con l'obiettivo di condensare in pochi termini il contributo scientifico dell'articolo, orientare la sua indicizzazione nei principali database bibliografici internazionali (come *Scopus* e *Web of Science*), e aumentarne la visibilità nei motori di ricerca accademici. Oltre a svolgere una funzione di sintesi semantica del contenuto, le parole chiave agiscono quindi come vettori informativi strategici, in grado di influenzare significativamente la probabilità che un articolo venga letto, scaricato o citato. In quest'ottica, l'approccio ITS assume che le frasi che contengono uno o più termini chiave siano maggiormente rappresentative del contenuto scientifico dell'articolo. Per tener conto di questa ipotesi, ITS pesa in modo differenziato le frasi in base alla presenza e alla densità di parole chiave al loro interno: le frasi che ne contengono di più ricevono un peso maggiore nella costruzione del grafo, aumentando la loro probabilità di essere selezionate come salienti. Di conseguenza, l'ITS non si limita a

valutare le frasi sulla base della sola similarità lessicale o posizione testuale, ma integra anche un criterio semantico guidato dal contenuto e coerente con le intenzioni dichiarative degli autori, ottenendo una sintesi maggiormente allineata con i nuclei concettuali dell'articolo.

Per rafforzare ulteriormente il ruolo semantico delle parole chiave nel processo di sintesi, l'ITS espande automaticamente il set iniziale di keyword fornite dagli autori attraverso l'integrazione di tecniche di estrazione automatica. Questa espansione consente di arricchire la rappresentazione concettuale del documento, combinando parole chiave esplicite, deliberate e strategiche, con parole chiave implicite, ovvero termini emergenti dal contenuto testuale che ne riflettono le strutture tematiche latenti. A tal fine, ITS impiega due algoritmi complementari: RAKE (*Rapid Automatic Keyword Extraction*; Rose *et al*., 2010) e TextRank per la *keyword extraction* (Mihalcea & Tarau, 2004). Il primo, RAKE, è un algoritmo non supervisionato basato sull'analisi delle co-occorrenze tra parole contigue, ignorando *stopword* e punteggiatura. Il testo viene segmentato in frasi e successivamente in sottostringhe composte da termini significativi (cioè privi di funzioni grammaticali). Ogni parola riceve un punteggio calcolato in base al numero di co-occorrenze con altre parole e alla frequenza con cui compare nei vari contesti. I termini candidati come parole chiave sono infine ottenuti sommando i punteggi delle parole che li compongono. RAKE è particolarmente efficace nel catturare espressioni multi-termine e unità lessicali specifiche di dominio. Il secondo metodo, TextRank, adotta una logica simile a quella impiegata per la selezione di frasi, ma applicata a livello lessicale. In questo caso, le singole parole significative del testo (escludendo le *stopword*) vengono rappresentate come nodi all'interno di un grafo, i cui archi connettono parole che co-occorrono entro una finestra scorrevole di contesto di dimensione prefissata (ad esempio, 2 o 3 parole). A questo grafo viene applicato l'algoritmo PageRank, che assegna a ciascun nodo un punteggio di centralità sulla base della sua posizione e connessioni nella rete. I termini con i punteggi di centralità più elevati vengono selezionati come parole chiave candidate. Un elemento distintivo di questa tecnica è che, se due termini adiacenti identificati come parole chiave compaiono consecutivamente all'interno di una frase, essi sono combinati in un'unica espressione multi-termine, riflettendo la presenza di unità lessicali composte (ad esempio, *social media*, *climate change*). In tal modo, l'algoritmo riesce a cogliere non solo i termini salienti individuali, ma anche collocazioni semantiche ricorrenti, migliorando l'espressività e la precisione del set finale di keyword. Questo approccio si dimostra particolarmente efficace nell'individuare termini centrali e ben connessi nel tessuto linguistico del testo, offrendo una rappresentazione sintetica e coerente del contenuto.

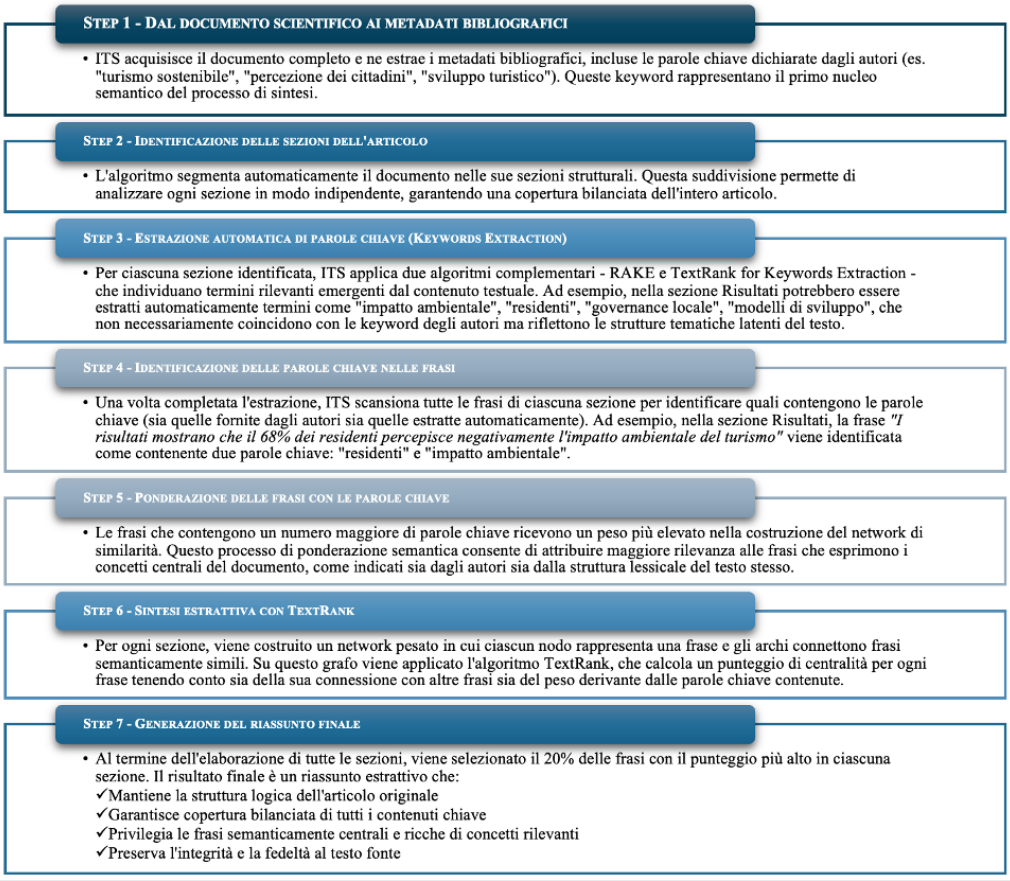
Combinando i due algoritmi di estrazione, l'ITS genera un insieme ampliato e articolato di parole chiave, in grado di rappresentare sia le dimensioni semantiche esplicite

sia quelle implicite del testo scientifico. Questo set viene quindi utilizzato per ponderare la rilevanza delle frasi, attribuendo maggiore peso a quelle che contengono un numero più elevato di keyword.

Il processo viene applicato iterativamente a ciascuna sezione del documento. In ogni sezione, l'ITS costruisce un grafo pesato delle frasi, nel quale i nodi sono influenzati sia dalla loro similarità testuale sia dalla densità di termini strategici. L'algoritmo TextRank viene quindi impiegato per classificare le frasi in base alla loro centralità nella rete. Al termine di ciascuna iterazione per sezioni, viene selezionato il 20% delle frasi più rilevanti, costituendo così un riassunto che riflette in modo bilanciato l'intero arco informativo del documento. Per chiarire il funzionamento dell'approccio ITS, un esempio semplificato applicato a un ipotetico articolo scientifico sulla sostenibilità del turismo è descritto nella Figura 2, seguendo i passaggi illustrati nel framework metodologico.

L'intera strategia dell'ITS e tutte le relative analisi sono state realizzate all'interno dell'ambiente R (versione 4.5.0). Per l'implementazione sono stati utilizzati i pacchetti textrank e udpipe per la costruzione del grafo e l'estrazione RAKE, rispettivamente. Algoritmi specifici sono stati sviluppati per l'identificazione iterativa delle parole chiave e il conteggio delle loro occorrenze nelle frasi, secondo una logica completamente automatizzata.

Figura 1. Esempio illustrativo del funzionamento dell'ITS



Fonte: Elaborazione degli Autori

4. Esperimento comparativo: ITS, TextRank e ChatGPT

L'efficacia dell'approccio ITS è stata valutata su un campione di dieci articoli scientifici appartenenti a discipline diverse, con particolare attenzione ai settori della sanità e delle scienze sociali (Tabella 1). La selezione dei documenti è stata progettata con due obiettivi metodologici ben definiti: da un lato, verificare la robustezza e generalizzabilità dell'algoritmo su testi provenienti da ambiti disciplinari eterogenei; dall'altro, costruire benchmark qualitativo basato sul contributo diretto di esperti del dominio.

Per ciascun articolo è stato contattato un autore, chiedendo di identificare, senza ricevere alcuna informazione sulle finalità del progetto, le frasi ritenute più rilevanti, con l'unico vincolo di selezionarne almeno una per ciascuna sezione del proprio contributo. Questo approccio ha consentito di raccogliere un insieme di annotazioni indipendenti e affidabili, che riflettono il giudizio esperto sull'importanza semantica delle frasi nel

contesto del singolo documento. Tali annotazioni sono state successivamente utilizzate come benchmark di riferimento per valutare le prestazioni dell'algoritmo ITS, permettendo un confronto diretto tra le frasi selezionate automaticamente e quelle indicate dagli autori, secondo una logica di validazione basata su corrispondenza semantica e copertura contenutistica.

Tabella 1. Campione di articoli scientifici selezionati per l'analisi

ID	*Riferimento*	*Keywords degli Autori*	*Numero di sezioni*	*Frasi per sezioni Mean ± SD*
doc_01	D'Aniello *et al.* (2022)	Academic Health Centers; Health policy; Healthcare configurations; Scientific productivity; Research impact	11	18.8 ± 11.56
doc_02	Aria *et al.* (2023)	Tourism impact; Citizens' perceptions; Tourism development; Tourism sustainability; Structural equation models	9	26.8 ± 21.52
doc_03	Robinson *et al.* (2016)	Data Citation Index; Data sharing; Citation practices; Scholarly communication; Repositories	7	27 ± 13.33
doc_04	Robinson *et al.* (2014)	Altmetric.com; Twitter; Mendeley; altmetrics; social impact; coverage; Web 2.0	5	21 ± 11.47
doc_05	Aria *et al.* (2020)	Quality of life; Bibliometric analysis; Thematic analysis	7	53 ± 21.14
doc_06	Aria *et al.* (2022)	Text analytics; Topic detection; Thematic mapping	7	26.43 ± 9.03
doc_07	Della Corte *et al.* (2018)	Destination governance; Distrust; Trust	7	46.7 ± 21.52
doc_08	D'Aniello *et al.* (2018)	Dogs; Human emotional smell; Interspecies emotional transfer; Emotional communication; Dog's heart rate; Dog–human bond	15	12.9 ± 13.68
doc_09	Ciavolino *et al.* (2022)	Partial least squares; Structural equation modelling; PLS-SEM; Bibliometrics citation analysis; Bibliometrix R package	13	23 ± 21.83
doc_10	Adamo *et al.* (2023)	Keratotic oral lichen planus; Depression; Anxiety; Mood disorder; Pain	9	13.8 ± 14.45

Fonte: elaborazioni degli Autori

Ai fini dell'esperimento, i testi completi degli articoli selezionati sono stati estratti e organizzati in data frame, con un file distinto per ciascun documento. Per garantire un'analisi focalizzata sui contenuti scientifici sostanziali, sono state escluse le sezioni marginali rispetto al corpo argomentativo del testo, quali l'abstract, i ringraziamenti, i materiali supplementari e la bibliografia. L'elaborazione ha quindi riguardato esclusivamente le sezioni comprese tra l'introduzione e la discussione / conclusione.

Una volta acquisiti, i testi sono stati importati nell'ambiente R per essere sottoposti a una fase preliminare di pre-processing linguistico, necessaria per la successiva applicazione degli algoritmi di estrazione. In particolare, il pre-processing ha incluso tre passaggi fondamentali:

1. *Tokenizzazione*, ovvero la suddivisione del testo in unità lessicali elementari (token), corrispondenti a parole o simboli discreti.
2. *Lemmatizzazione*, ovvero la riduzione dei token alla loro forma canonica (lemma), mediante la rimozione di prefissi, suffissi e varianti flessionali, così da uniformare le forme lessicali e favorire una rappresentazione coerente e comparabile del contenuto semantico (ad esempio, le parole "corre", "correva", "correndo" vengono tutte ricondotte al lemma "correre").
3. *Part-of-Speech (PoS) tagging*, ossia l'attribuzione di un'etichetta grammaticale a ciascun token (ad esempio, sostantivo, verbo, aggettivo), che definisce il suo ruolo sintattico.

A valle di queste operazioni, è stato possibile identificare ed etichettare i termini corrispondenti alle parole chiave, sia quelle dichiarate dagli autori, sia quelle individuate automaticamente tramite gli algoritmi RAKE e TextRank. Tali termini sono stati taggati come "keyword" e mantenuti integralmente nelle fasi successive, allo scopo di attribuire maggiore peso alle frasi che li contengono, secondo la logica illustrata nei paragrafi precedenti. Questo passaggio ha permesso di integrare l'informazione semantica nel processo di selezione delle frasi più rilevanti, rafforzando la capacità dell'algoritmo ITS di cogliere i nuclei concettuali del testo.

Completata la fase di pre-processing linguistico e di identificazione delle parole chiave, l'algoritmo ITS è stato applicato iterativamente a ciascuna sezione degli articoli per procedere con l'estrazione delle frasi ritenute più rilevanti e selezionarne, quindi, il 20% con il punteggio più alto, generando un riassunto sintetico ma rappresentativo della struttura logica del documento.

Per valutare le performance dell'approccio ITS, si è proceduto al confronto tra le frasi selezionate automaticamente e quelle indicate dagli autori degli articoli, utilizzate come benchmark di riferimento. In particolare, è stato conteggiato il numero di corrispondenze tra le frasi individuate da ITS e quelle segnalate dagli autori, al fine di

misurare la capacità dell'algoritmo di intercettare i contenuti ritenuti centrali da esperti del dominio.

A scopo comparativo, è stata adottata come *baseline* l'implementazione classica dell'algoritmo TextRank, priva di ponderazione semantica tramite parole chiave. Questo confronto ha consentito di isolare il contributo specifico dell'integrazione delle keyword nel miglioramento delle prestazioni.

Infine, è stato incluso nell'analisi comparativa anche il sistema ChatGPT (modello GPT-4o), data la sua crescente adozione nel contesto accademico per compiti di sintesi automatica, recupero delle informazioni e riscrittura stilistica. A ciascun documento è stato applicato un prompt standardizzato, volto a richiedere l'estrazione delle frasi più rilevanti per ogni sezione, escludendo abstract, ringraziamenti e bibliografia. L'inclusione di ChatGPT ha permesso di posizionare l'algoritmo ITS anche in relazione a un modello linguistico avanzato di intelligenza artificiale, valutandone le prestazioni in un contesto di confronto multilivello. Per ottenere l'estrazione delle frasi rilevanti dai testi scientifici tramite ChatGPT, è stato caricato il pdf di ogni documento e usato il seguente prompt:

> *"Extract the 20% most relevant sentences for understanding the content of the attached scientific article for each section and subsection of the document. Exclude sections such as the abstract, acknowledgment, supplementary data, and bibliography"*.

4.1 Risultati

La Tabella 2 presenta un confronto sistematico tra i tre metodi di sintesi automatica considerati – TextRank, ITS e GPT-4o – in termini di capacità di identificare correttamente le frasi rilevanti negli articoli del campione. Per ciascun documento, sono stati calcolati il numero e la percentuale di frasi estratte automaticamente che corrispondono a quelle annotate dagli autori come più significative. Questa metrica permette di valutare in che misura ogni algoritmo riesce a catturare i contenuti essenziali secondo il giudizio di esperti del dominio. Osservando i dati in tabella, emergono pattern interessanti. ITS raggiunge in media 12.5 frasi correttamente identificate per documento (mediana, con un intervallo interquartile IQR: 9-15.5), superando sia TextRank (mediana: 8.5 frasi, IQR: 6-11.8) sia GPT-4o (mediana: 5 frasi, IQR: 3.25-7.75).

In diversi documenti, l'ITS ha dimostrato capacità superiori. Nel documento doc_07 (Della Corte et al., 2018), l'ITS identifica correttamente 16 frasi su 25 annotate dagli autori (64%), superando nettamente TextRank (11 frasi, 44%) e soprattutto GPT-4o (solo 3 frasi, 12%). Analogamente, nel documento doc_05 (Aria et al., 2020), l'ITS raggiunge un tasso di accuratezza del 56.7% (17 frasi su 30), mentre GPT-4o si ferma al 26.7% e TextRank al 43.3%. Questi risultati suggeriscono che l'integrazione delle keyword nel

processo di ponderazione delle frasi migliora significativamente la capacità di identificare contenuti centrali, soprattutto in articoli con keyword ben distribuite nel testo.

Le performance del modello generativo mostrano la maggiore variabilità. In alcuni casi, GPT-4o ha ottenuto risultati eccellenti: nel documento doc_01 (D'Aniello *et al*., 2022) identifica correttamente 22 frasi su 35 (62.9%), superando sia l'ITS (19 frasi, 54.3%) sia il TextRank (12 frasi, 34.3%). Questo risultato testimonia le potenti capacità di comprensione semantica dei LLMs. Tuttavia, in altri documenti si osservano crolli improvvisi: doc_07 (12%), doc_09 (11.1%), doc_10 (16.7%). L'analisi qualitativa ha rivelato che GPT-4o tende a generare frasi parafrasate o sintetiche che, pur coerenti semanticamente, non corrispondono a passaggi testuali effettivi dell'articolo originale, compromettendo così l'allineamento con le annotazioni degli autori che si riferiscono a frasi specifiche del testo.

Per verificare se le differenze osservate tra i tre metodi fossero statisticamente significative e non dovute al caso, è stato applicato il test non parametrico di Friedman ($\chi^2 = 10.3$, gradi di libertà = 2, p-value = 0.006**), appropriato per confronti su misure ripetute in campioni di dimensione ridotta. Il test ha confermato l'esistenza di differenze significative tra i metodi. Per identificare quali coppie di metodi differissero in modo rilevante, è stato quindi condotto un test *post-hoc* di Durbin-Conover, i cui risultati sono riportati nella Tabella 3.

Tabella 2. Accuratezza dell'identificazione delle frasi rilevanti: confronto tra TextRank, ITS e GPT-4o

ID	*Frasi annotate dagli autori N*	*Frasi identificate con il TextRank N (%)*	*Frasi identificate con ITS N (%)*	*Frasi identificate con GPT-4o N (%)*
doc_01	35	12 (34.3%)	19 (54.3%)	**22 (62.9%)**
doc_02	42	13 (31%)	**14 (33.3%)**	12 (28.6%)
doc_03	17	6 (35.3%)	**7 (41.2%)**	**7 (41.2%)**
doc_04	8	**3 (37.5%)**	**3 (37.5%)**	**3 (37.5%)**
doc_05	30	13 (43.3%)	**17 (56.7%)**	8 (26.7%)
doc_06	21	8 (38.1%)	**12 (57.1%)**	3 (14.3%)
doc_07	25	11 (44%)	**16 (64%)**	3 (12%)
doc_08	33	9 (27.3%)	**13 (39.4%)**	5 (15.2%)
doc_09	36	5 (13.9%)	**12 (33.3%)**	4 (11.1%)
doc_10	30	6 (20%)	**8 (26.7%)**	5 (16.7%)
Mediana [IQR]	30 [22-34.5]	8.5 [6-11.8]	12.5 [9-15.5]	5 [3.25-7.75]

Fonte: elaborazioni degli Autori

Tabella 3. Confronto statistico tra TextRank, ITS e GPT-4o: risultati dei test di Friedman e Durbin-Conover

Friedman test	**χ^2**	**df**	**p-value**
10.3	2	**0.006****	
Durbin-Conover test: Pairwise comparison		**Statistics**	**p-value**
ITS - TextRank		3.10	**0.006****
TextRank - GPT-4o		1.14	0.268
ITS - GPT-4o		4.24	**<0.001****

*Livelli di significatività del p-value: *0.01 < p-value ≤ 0.05; **p-value ≤ 0.01*

Fonte: elaborazioni degli Autori

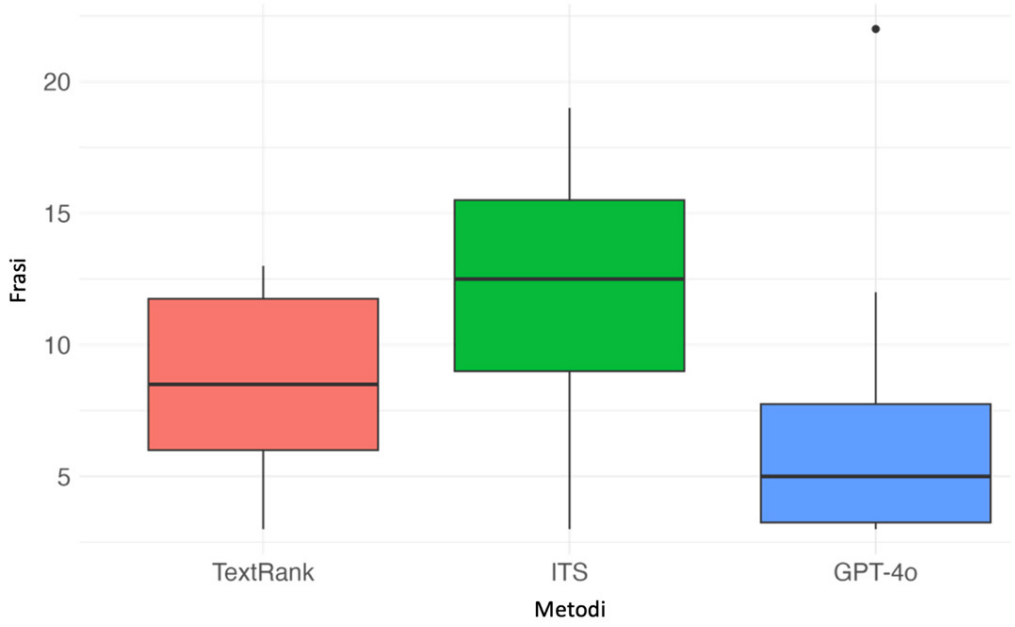
I risultati dell'analisi statistica (Tabella 3) forniscono un quadro chiaro delle relazioni tra i metodi:

- **ITS vs TextRank:** La differenza è statisticamente significativa (p-value = 0.006**), confermando che l'integrazione delle parole chiave apporta un contributo sostanziale rispetto all'approccio classico basato sulla sola similarità lessicale.
- **ITS vs GPT-4o:** Anche in questo caso la differenza è altamente significativa (p-value < 0.001**), evidenziando che, nonostante GPT-4o rappresenti un modello linguistico di frontiera, nella specifica task di sintesi estrattiva di documenti scientifici completi le sue performance complessive sono inferiori a quelle di ITS.
- **TextRank vs GPT-4o:** Sorprendentemente, questi due metodi non presentano differenze statisticamente significative (p-value = 0.268). Questo risultato indica che, mediamente, un algoritmo classico come TextRank si comporta in modo comparabile a un modello generativo avanzato come GPT-4o nel compito di identificare frasi rilevanti in articoli scientifici, suggerendo che la complessità computazionale dei LLM non si traduce necessariamente in vantaggi prestazionali in questo contesto specifico.

La rappresentazione grafica della Figura 2 visualizza efficacemente le differenze tra i tre metodi attraverso box plot che sintetizzano la distribuzione del numero di frasi correttamente identificate per ciascun approccio. ITS mostra non solo una mediana superiore, ma anche una distribuzione complessivamente più elevata rispetto agli altri metodi. La presenza di valori che si estendono fino a 19 frasi corrette testimonia la capacità dell'algoritmo di raggiungere performance eccellenti in diversi documenti. TextRank

evidenzia una distribuzione più contenuta e simmetrica, con valori concentrati tra 6 e 12 frasi, indicando prestazioni mediamente stabili ma inferiori a ITS. GPT-4o presenta la maggiore variabilità: pur mostrando alcuni valori estremi superiori (picco a 22 frasi nel doc_01), la sua mediana è la più bassa dei tre metodi e il box è concentrato su valori bassi (tra 3 e 8 frasi). Questo pattern visivo conferma quanto emerso dall'analisi numerica: GPT-4o alterna casi di eccellenza a fallimenti sostanziali, risultando complessivamente meno affidabile per il compito di sintesi estrattiva di documenti scientifici completi. Il grafico rafforza dunque le evidenze statistiche, sottolineando come ITS offra il miglior compromesso tra accuratezza media e stabilità delle performance tra documenti diversi.

Figura 2. Distribuzione delle frasi identificate da TextRank, ITS e GPT-4o



Fonte: elaborazione degli Autori

5. Conclusioni e discussione

I risultati emersi dall'applicazione dell'ITS mettono in evidenza l'importanza di integrare informazioni strutturali e semantiche nei processi di sintesi automatica dei testi scientifici. L'approccio ITS, grazie alla sua architettura basata sull'analisi sezionale e sull'uso strategico delle parole chiave (sia fornite dagli autori sia estratte automaticamente), si è dimostrato efficace nel selezionare frasi rilevanti, spesso in coerenza con le annotazioni degli autori. Anche nei casi in cui non vi era una corrispondenza esatta,

le frasi individuate offrivano comunque contributi informativi rilevanti, arricchendo la comprensione dei documenti analizzati.

A confronto con altri approcci, ITS si è distinto non solo per accuratezza, ma anche per coerenza delle performance tra articoli e sezioni. In particolare, l'algoritmo ha superato significativamente sia la versione standard di TextRank sia il modello linguistico avanzato GPT-4o, soprattutto in termini di affidabilità nella selezione delle frasi e stabilità dei risultati.

L'analisi ha infatti evidenziato limiti strutturali nei modelli linguistici di grandi dimensioni (Large Language Model - LLM), come GPT-4o, in particolare nella gestione di testi lunghi, quali gli articoli scientifici completi. Nonostante la capacità teorica di elaborare input estesi (fino a 4.000 token), i documenti analizzati in questo studio superavano spesso tale soglia, determinando un comportamento incoerente del modello. In alcuni casi, GPT-4o ha generato frasi parafrasate o addirittura non presenti nel testo originale, compromettendo l'aderenza al contenuto e la validità della sintesi.

A differenza degli LLM, ITS non presenta vincoli di lunghezza testuale, poiché elabora il documento in modo iterativo e sezione per sezione, mantenendo sempre un controllo diretto sull'origine e sulla selezione delle frasi. Questo rende l'approccio non solo più trasparente, ma anche più interpretabile**,** qualità essenziali in contesti come quello accademico, dove è fondamentale comprendere e giustificare i criteri con cui le informazioni vengono selezionate.

Ciò nonostante, è importante riconoscere che i modelli generativi, come GPT-4o, eccellono nella velocità di sintesi e nella capacità di produrre testi coesi e contestualmente ricchi, offrendo un supporto utile per l'analisi esplorativa e il reperimento rapido delle informazioni. La loro efficacia nella riduzione della mole testuale può rappresentare un vantaggio significativo per i ricercatori, soprattutto nelle prime fasi di screening bibliografico. Tuttavia, le criticità in termini di precisione, coerenza e tracciabilità delle fonti impongono cautela nel loro impiego come strumenti autonomi di estrazione di conoscenza.

A livello epistemologico, questi risultati sollevano interrogativi più ampi circa il ruolo dell'intelligenza artificiale nella ricerca scientifica. Se da un lato i LLM contribuiscono a democratizzare l'accesso al sapere, dall'altro pongono il problema della fiducia e dell'affidabilità: quanto possiamo affidarci a modelli che, seppur potenti, possono generare contenuti inesatti o inventati? E quale impatto avrà la crescente diffusione di tali strumenti sulla profondità dell'analisi critica e sull'interazione diretta con i testi originali?

In questo scenario, approcci come ITS, che combinano efficienza computazionale, trasparenza metodologica e rigore scientifico, rappresentano una risposta promettente, in grado di coniugare l'automazione con l'esigenza di controllo interpretativo. La sin-

tesi automatica non sostituisce l'interpretazione umana, ma può agire come strumento complementare per supportare, e non soppiantare, il pensiero critico nella gestione della conoscenza.

In conclusione, se da un lato i modelli di linguaggio su larga scala costituiscono un potente alleato nel fronteggiare l'overload informativo, dall'altro è necessario continuare a sviluppare strumenti più trasparenti, controllabili e adattabili ai requisiti della ricerca scientifica. ITS si colloca in questa direzione, offrendo una soluzione solida per la sintesi di documenti complessi e confermandosi un valido supporto alla comprensione e all'analisi della letteratura scientifica, in un contesto sempre più dominato dall'intelligenza artificiale, ma che richiede ancora rigore metodologico e garanzie di affidabilità.

6. Sviluppi futuri e limitazioni

Il presente studio presenta alcune limitazioni che aprono interessanti prospettive di ricerca futura. In primo luogo, il campione analizzato, sebbene multidisciplinare, è limitato a dieci articoli scientifici. Studi futuri potrebbero estendere la valutazione a corpus più ampi e diversificati per disciplina, lingua e tipologia di pubblicazione (review sistematiche, meta-analisi, studi empirici), al fine di verificare e rafforzare la generalizzabilità dell'approccio ITS in contesti più eterogenei.

Inoltre, il confronto con GPT-4o, condotto utilizzando il modello disponibile al momento delle analisi (2024), sarà essere aggiornato considerando l'evoluzione continua dei LLMs. L'ecosistema dell'intelligenza artificiale è in rapida trasformazione, con il rilascio frequente di nuovi modelli non solo da parte di OpenAI (GPT-5 e versioni successive), ma anche di altri provider come Google (Gemini), Anthropic (Claude), e Meta (LLaMA). Ciascuno di questi sistemi presenta architetture, capacità e limitazioni specifiche che meritano un'analisi comparativa approfondita.

Tuttavia, è importante sottolineare che le criticità osservate in GPT-4o, in particolare la gestione problematica di documenti lunghi, la tendenza a generare contenuti parafrasati o non presenti nel testo originale, e l'instabilità nelle performance, non sono esclusivamente legate alla versione specifica del modello, ma riflettono caratteristiche strutturali comuni ai LLM generativi. Anche modelli più avanzati, pur migliorando in termini di coerenza e verosimiglianza delle risposte, continuano a presentare sfide in termini di:

- Tracciabilità: difficoltà nel garantire che ogni affermazione possa essere ricondotta con certezza a una specifica porzione del testo originale.
- Fedeltà semantica: rischio di introdurre interpretazioni o riformulazioni che, pur plausibili, possono discostarsi dal significato originario.
- Riproducibilità: variabilità intrinseca nelle risposte generate a parità di input, che può compromettere la stabilità dei risultati in applicazioni scientifiche.

Al contrario, l'approccio ITS offre garanzie strutturali di trasparenza, interpretabilità e aderenza al testo fonte, caratteristiche che rimangono rilevanti indipendentemente dai progressi tecnologici nei modelli generativi. Pertanto, piuttosto che considerare ITS e i LLM come approcci in competizione, appare più produttivo esplorarne le potenziali sinergie: i modelli generativi potrebbero essere impiegati per compiti esplorativi e di prima analisi, mentre metodi estrattivi come ITS potrebbero garantire rigore e verificabilità nella selezione finale dei contenuti.

Dal punto di vista applicativo, ITS potrebbe essere integrato in piattaforme di gestione bibliografica (come Zotero, Mendeley, o EndNote) per fornire riassunti automatici di articoli importati, facilitando le fasi di screening preliminare nelle revisioni sistematiche della letteratura. Un'ulteriore applicazione riguarda il supporto alla stesura di literature reviews: l'algoritmo potrebbe essere impiegato per estrarre automaticamente le frasi più rilevanti da insiemi di articoli correlati, permettendo ai ricercatori di confrontare rapidamente contributi, metodologie e risultati empirici senza dover leggere integralmente decine di documenti. Infine, in ambito editoriale e di peer review, l'approccio ITS potrebbe assistere revisori ed editor nella rapida valutazione della coerenza strutturale di un manoscritto e nell'identificazione dei contributi chiave dichiarati dagli autori, riducendo i tempi di prima valutazione.

Ringraziamenti

La presente ricerca è stata realizzata con il supporto dei seguenti progetti finanziati nell'ambito del PRIN 2022:

(1) SCIK-HEALTH (Codice Progetto: 2022825Y5E – CUP: E53D23006110006);

(2) PNRR – The value of scientific production for patient care in Academic Health Science Centres (Codice Progetto: P2022RF38Y – CUP: E53D23016650001).

Bibliografia

Adamo, D., Calabria, E., Canfora, F., Coppola, N., Leuci, S., Mignogna, M., Lo Muzio, L. et al. (2023). Anxiety and depression in keratotic oral lichen planus: a multicentric study from the SIPMO. *Clinical Oral Investigations, 27*(6), 3057-3069. 10.1007/s00784-023-04909-3.

Aria, M., Cuccurullo, C., D'Aniello, L., Misuraca, M., & Spano, M. (2022). Text summarization of a scientific document: a comparison of extractive unsupervised methods. In *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data* (Vol. 1, pp. 67-73). Napoli: VADISTAT Press/Edizioni Erranti.

Aria, M., Cuccurullo, C., D'Aniello, L., Misuraca, M., & Spano, M. (2022). Thematic analysis as a new culturomic tool: the social media coverage on COVID-19 pandemic in Italy. *Sustai-*

nability, 14(6), 3643. https://doi.org/10.3390/su14063643.

Aria, M., D'Aniello, L., Della Corte, V., & Pagliara, F. (2023). Balancing tourism and conservation: analysing the sustainability of tourism in the city of Naples through citizen perspectives. *Quality & Quantity, 58*, 1-21. https://doi.org/10.1007/s11135-023-01774-w.

Aria, M., Misuraca, M., & Spano, M. (2020). Mapping the evolution of social research and data science on 30 years of social indicators research. *Social Indicators Research, 149*, 803-831. https://doi.org/10.1007/s11205-020-02281-3.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901. Retrieved from https://arxiv.org/abs/2005.14165.

Chen, Y. C., & Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. arXiv preprint arXiv:1805.11080. Retrieved from https://arxiv.org/abs/1805.11080.

Ciavolino, E., Aria, M., Cheah, J. H., & Roldán, J. L. (2022). A tale of PLS structural equation modelling: episode I-a bibliometrix citation analysis. *Social Indicators Research, 164*(3), 1323–1348. https://doi.org/10.1007/s11205-022-02994-7.

D'Aniello, B., Semin, G. R., Alterisio, A., Aria, M., & Scandurra, A. (2018). Interspecies transmission of emotional information via chemosignals: from humans to dogs (Canis lupus familiaris). *Animal Cognition, 21*, 67-78. https://doi.org/10.1007/s10071-017-1139-x.

D'Aniello, L., Spano, M., Cuccurullo, C., & Aria, M. (2022). Academic Health Centers' configurations, scientific productivity, and impact: Insights from the Italian setting. *Health Policy, 126*(12), 1317-1323. https://doi.org/10.1016/j.healthpol.2022.09.007.

D'Aniello, L., Aria, M., Cuccurullo, C., Misuraca, M., & Spano, M. (2024). Extracting knowledge from scientific literature with an integrated Text Summarization approach. In A. Dister & D. Longrée (Eds.), *Mots competes textes déchiffrés* (Vol. 1, pp. 239-248). Louvain: Presses Universitaires De Louvain.

Della Corte, V., Aria, M., & Del Gaudio, G. (2018). Strategic governance in tourist destinations. *International Journal of Tourism Research, 20*(4), 411-423. https://doi.org/10.1002/jtr.2192.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457-479. https://doi.org/10.1613/jair.1523.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review, 47*, 1-66. https://doi.org/10.1007/s10462-016-9475-9.

Koh, H. Y., Ju, J., Liu, M., & Pan, S. (2022). An empirical survey on long document summarization: Datasets, models and metrics. *ACM Computing Surveys, 55*(8), 1-35. https://doi.org/10.1145/3545176.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature, 535*(7612), 457-458. https://doi.org/10.1038/nj7612-457a.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer,

L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. Retrieved from https://arxiv.org/abs/1910.13461.

Mani, I. (2001). *Automatic summarization*. Amsterdam: John Benjamins Publishing.

Meng, R., Thaker, K., Zhang, L., Dong, Y., Yuan, X., Wang, T., & He, D. (2021). Bringing structure into summaries: a faceted summarization dataset for long scientific documents. arXiv preprint arXiv:2106.00130. Retrieved from https://arxiv.org/abs/2106.00130.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404-411). Barcelona: Association for Computational Linguistics.

Nenkova, A., & McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval, 5*(2-3), 103-233. https://doi.org/10.1561/1500000015.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: Exploring the insides of Altmetric.com. *El Profesional de la Información, 23*(4), 359–366. https://doi.org/10.3145/epi.2014.jul.03.

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology, 67*(12), 2964-2975. https://doi.org/10.1002/asi.23529.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. In M. W. Berry & J. Kogan (Eds.), *Text Mining: Applications and Theory* (pp. 1-20). Chichester: John Wiley & Sons.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., et al. (2017). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics, 54*, 202-212. https://doi.org/10.1016/j.jbi.2015.02.004.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., et al. (2020). Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems, 33*, 17283-17297. Retrieved from https://arxiv.org/abs/2007.14062.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 11328-11339). PMLR. Retrieved from http://proceedings.mlr.press/v119/zhang20ae.html.

APPROFONDIMENTI SU TWEET IN LINGUA ITALIANA: DISTRIBUZIONI, CONTENUTI, SENTIMENT, MULTIMEDIA E METRICHE DI RETE

INSIGHTS FROM ITALIAN TWEETS: DISTRIBUTIONS, CONTENT, SENTIMENT, MULTIMEDIA, AND NETWORK METRICS

Domenica Fioredistella Iezzi[1], *Roberto Monte*[2] *and Daniele Pasquini*[3]

Sommario

Comprendere i meccanismi che guidano la viralità consente di individuare i fattori che plasmano la popolarità e i livelli di coinvolgimento delle principali tendenze nei social su diversi domini. Questo lavoro persegue tre obiettivi principali: (1) analizzare la distribuzione dei messaggi su un ampio campione di tweet; (2) modellare i pattern temporali dei messaggi virali e identificare le principali macro-dimensioni che contribuiscono alla natura virale dei contenuti social; (3) esaminare la struttura di rete degli utenti che hanno pubblicato tali contenuti. La letteratura individua due principali approcci alla previsione delle condivisioni: la predizione della popolarità basata sul contenuto dei messaggi, analizzando i testi e i contenuti multimediali dei post, e quella della popolarità basata sulla struttura della rete. In questo ultimo approccio, si modella la struttura della rete per comprendere come i messaggi vengano condivisi tra gli utenti. La nostra ricerca adotta un approccio a metodi misti, integrando analisi del sentiment, indicatori chiave di performance e metriche di popolarità degli utenti, al fine di caratterizzare le componenti della viralità. Per testare il modello proposto, analizziamo un dataset di 22.155.362 tweet italiani, pubblicati tra il 1° dicembre 2020 e il 12 dicembre 2020.

Abstract

Understanding the mechanisms that drive virality can reveal the factors shaping the popularity and engagement levels of central societal trends and topics. This paper has three main objectives: (1) to analyze message distribution across a large sample of tweets, (2) to model the temporal patterns of viral messages and identify key macro-dimensions contributing to the viral nature of social content, and (3) to exa-

[1] Università di Tor Vergata, Department of Enterprise Engineering "Mario Lucertini", Rome, Italy - e-mail: stella.iezzi@uniroma2.it

[2] Università di Tor Vergata, Department of Civil Engineering and Computer Science Engineering, Rome, Italy - e-mail: roberto.monte@uniroma2.it

[3] Università di Tor Vergata, Department of Enterprise Engineering "Mario Lucertini", Rome, Italy - e-mail: psqdni@hotmail.it

mine the network structure of users who posted this content. The literature identifies two primary approaches to predicting shares: Content-based popularity prediction, which examines textual and multimedia attributes within posts, and Circulation-based popularity prediction, which models the network structure to understand how posts spread among users. Our research employs a mixed-method approach, integrating sentiment analysis, key performance indicators, and user popularity metrics to characterize the components of virality. To test our model, we analyze a dataset of 22,155,362 Italian tweets from December 1, 2020, to December 12, 2020.

Parole chiave: viralità nei social media, big data, rete sociale, analisi del sentiment, modello di regressione.
Keywords: *virality in social media, big data, social network, sentiment analysis, regression model.*

1. Introduction

The number of people using social media to share information is steadily increasing. Social media is a digital platform that connects individuals, enables content creation and sharing, facilitates knowledge exchange, and preserves valuable information for future access (Ghaisani *et al.,* 2017). According to We Are Social and Meltwater (2024), the number of active social media profiles worldwide has surpassed 5 billion, reaching 5.04 billion, over 62% of the global population. This global total has grown by 266 million in the past year, reflecting an annual increase of 5.6%. This remarkable figure shows that, over the last year, the world has averaged an astounding 8.4 new social media users per second. The habits of Italians align with those of other countries around the world. According to the research presented in the report on Italians, a significant amount of time is spent online for various reasons. The primary reasons people access social media are to stay informed about current events and to entertain themselves in their free time (47%), followed closely by the desire to keep in touch with friends and family (45%).

Additionally, there is an increase in the daily time spent on social media and the number of people who report watching video content (91%). This growth is primarily driven by content in the "comedy, memes, and viral videos" category (+3.7%). As a result, certain types of content (posts, tweets, messages, short videos) are particularly engaging and quickly shared with many users.

We talk about virality, which refers to the ability of content to spread rapidly and widely on the internet, often through shares and word of mouth. It is a common phenomenon on social media where a video, post, or meme can go viral, reaching a vast audience quickly. The adjective "viral" comes from the Latin word "poison". According to

Dimmock *et al*. (2016), the discovery of virus's dates to 1892 when Dmitrij Ivanovsky described a non-bacterial pathogen capable of infecting tobacco plants in a paper. Later, the Oxford Dictionary included "viral" among adjectives, defining it as a neologism to indicate something "that spreads particularly quickly and widely, especially through new communication media" or "that tends to spread extensively." In social media, viral diffusion refers to how content quickly spreads across a digital platform. In this case, the "viral" content is spontaneously shared by many people, exponentially increasing its visibility. Messages that are highly retweeted, e.g., are social indicators to evaluate the ability of content (eventually accompanied by video or photos) to spread quickly and widely across social networks and online platforms. This indicator reflects the level of interest, engagement, and social relevance a piece of content generates among users, as sharing is often driven by emotional reactions, timeliness, novelty, or the desire to express one's identity and values. Retweet very popular is, therefore, a social indicator that provides insights into the dynamics of idea dissemination and how specific content reflects and influences collective values, emotions, and interests.

We want to discover the probability distribution of these tweets to understand the characteristics that can make a message go highly retweeted on the platform.

This paper aims to characterize virality by analyzing the distribution of messages from a large sample of tweets. It will model the temporal distribution of viral messages, identify key macro-dimensions contributing to the viral nature of social content, and examine the network structure of individuals who have posted the content. Additionally, the study will investigate the propagation time of a tweet to further understand these dynamics.

The network structure of all social platforms provides information about users and how they are connected through a web of relationships. In this structure, users, with their ties and interactions, such as followers, friends, or direct connections, form a network-like structure (or "grid").

We tested our model on 22,155,362 Italian tweets on various topics collected between December 1 and December 12, 2020. Of these tweets, 6,281,784 received at least one retweet, while 15,873,578 did not. Using a big data dataset, we aimed to validate the model's effectiveness in predicting content popularity across various topics.

The structure of this paper is as follows: Section 2 examines the framework and research directions; Section 3 explores the time for a tweet to become highly shared; Section 4 analyzes sentiment, multimedia elements, and topics within the most retweeted messages in our sample; Section 5 presents our network metrics; Section 6 details regression models for overdispersal count responses and presents our findings; finally, Section 7 concludes with implications and directions for future research.

2. Framework and Research Directions

Various studies have examined the factors influencing the online diffusion of information and electronic word-of-mouth (e-wom) in social networks, highlighting how these processes are analogous to viral contagion mechanisms. Ngo *et al*. (2024) explore the complex relationships among various dimensions of e-wom information, including its credibility, usefulness, adoption, and attitudes toward it. They examine how these factors collectively influence online purchase intentions. Phelps *et al*. (2004) analyze findings from three studies investigating consumer motivations and responses to forwarding emails. The authors discuss the implications for target selection and message creation, providing valuable insights for advertising practitioners looking to implement viral marketing strategies. Additionally, they offer recommendations for future research focused on computer-mediated consumer-to-consumer interactions, highlighting areas of interest for academic researchers.

Several studies (e.g., Berger & Milkman, 2012; De Vries *et al*., 2012; Phelps *et al*., 2004; Kwak *et al*., 2010; Trilling *et al*., 2017) have analyzed the phenomenon of online popularity, which serves as a key indicator of social behavior in digital environments. A variety of metrics can be used to assess popularity, including:

1. Number of Shares: the frequency with which users redistribute the content across their networks.
2. Number of Views: the total number of times the content has been accessed or watched.
3. Engagement: the level of user interaction with the content, encompassing likes, comments, and shares.
4. Growth Rate: the speed at which the content accumulates views and shares over a given time.
5. Reach: the total number of unique users who have encountered the content, either directly or via sharing.

Kim (2018) investigates how social media virality metrics impact perceptions of message influence on oneself and others and intentions to take preventive actions. In this online experiment, participants viewed a Facebook post discussing a health risk, with variations in virality metrics such as the number of likes and shares. The findings reveal distinct effects associated with these metrics: high share counts significantly enhanced perceived message influence on oneself and others and increased intentions to engage in preventive behaviors. Elmas *et al*. (2023) suggested using the ground truth data provided by Twitter's "Viral Tweets" topic to review the current metrics and propose new metrics. Tiago *et al*., 2019 analyze the content promoting tourist destinations. The YouTube platform, which has video content, has proven engaging in this sector.

Bene (2017) addresses the issue of virality in political content messages on Facebook. The results showed that citizens are highly reactive to posts containing negative emotions, text- based posts, personal posts, and those that require action. Virality is mainly facilitated by memes, videos, harmful content, and mobilizing posts, as well as posts containing a request for sharing. Analyzing the most frequently e-mailed New York Times (NYT) articles, Berger and Milkman (2012) found that content virality correlates positively with its positivity and emotional impact, particularly for emotions such as anger, awe, and anxiety, while it is negatively correlated with sadness. Using a sample of German articles, Heimbach and Hinz (2016) replicated their study for the most e-mailed list of Germany's leading news magazine and expanded the analysis to include (1) three additional communication channels and (2) the non-linear relationship between positivity and virality. From a methodological perspective, Avalle *et al*. (2024) provide a large-scale comparative analysis of online conversations across eight social media platforms and over three decades, encompassing more than 500 million comments. Their results demonstrate the persistence of heavy-tailed engagement distributions and invariant toxicity patterns, suggesting that platform design plays a secondary role compared to stable human behavioral dynamics. While we could not replicate BM's findings, our results align with their conclusions across all communication channels. Additionally, we propose that the relationship between positivity and virality exhibits an inverted U-shape, indicating a non-linear pattern.

Our research questions are as follows:

1. How can we characterize virality by examining the message distribution within a large sample of tweets?
2. What are the key macro-dimensions that contribute to the viral nature of social content when studying the temporal distribution of viral messages?
3. How does the network structure of individuals influence the diffusion of highly retweeted content? The probability distribution of retweets offers a comprehensive perspective on message propagation, enabling prediction, optimization, risk management, and informed decision-making.

3. Retweet distribution

To characterize the retweet distribution, we treat the number of retweets received by a single tweet as a count of failures: a tweet with zero retweets is considered to have zero failures, one retweet corresponds to one failure, and so on. This mirrors the conceptual framework of the geometric distribution, which models the number of failures before the first success with probability mass function (PMF) given by:

$$f(n|p) := \mathrm{p}(1-\mathrm{p})^{\mathrm{n}} \quad \mathrm{n} \in \mathbb{N}_0,\ \mathrm{p} \in (0,1),$$

where n is the number of failures before the first success and p is the success probability parameter. On the other hand, the tweets with zero failures vastly outnumber all other tweets. To capture this over-representation of zeros, we adopt a zero-inflated geometric (ZIG) distribution with PMF given by:

$$f(n|\varphi,p) := \begin{cases} \varphi + (1-\varphi)p, & if\ n = 0, \\ (1-\varphi)p(1-p)^n, & if\ n \in \mathbb{N}, \end{cases} \varphi, p \in (0,1),$$

which introduces an inflation parameter φ to increase the probability mass at $n = 0$. Another issue is that each tweet is unique in several respects, and it is unrealistic to assume a constant probability parameter p across all tweets. Instead, we treat p as a random variable, varying across tweets. This approach leads us to model the retweet distribution as a mixture of geometric distributions with different success probabilities. Formally, we assume that our PMF is given by

$$p(n|\vartheta) := \int_0^1 f(n|\varphi,p)\mathrm{g}(p|\vartheta)dp,$$

where $\mathrm{g}(p|\vartheta)$, referred to as mixing distribution, is the density of the success parameter p. Following the structure of "Buy 'Till You Die" (BTYD) models (see Ping *et al*., 2022; Chou *et al*., 2022), we model the mixing distribution of p using a beta density with shape parameters $\alpha, \beta > 0$, that is:

$$\mathrm{g}(p|\vartheta) \equiv \mathrm{g}(p|\alpha,\beta) := \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)},$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function. Consequently, a direct computation shows that our candidate distribution for the retweets results in a zero-inflated beta-geometric (ZIBG) distribution given by:

$$f(n|\varphi,\alpha,\beta) := \begin{cases} \varphi + (1-\varphi)\dfrac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)}, & if\ n = 0, \\ (1-\varphi)\dfrac{\mathrm{B}(\alpha+1,\beta+n)}{\mathrm{B}(\alpha,\beta)}, & if\ n \in \mathbb{N}, \end{cases} \varphi \in (0,1), \alpha, \beta > 0,$$

which offers a great flexibility in capturing the heterogeneity of tweet performance. Another advantage of the choice of the beta distribution as the mixing distribution is that it is possible to compute its first three moments in a closed form. This allows us

the application of the method of the moments to determine preliminary estimates of the parameters φ, α, β, through computational procedures for solving nonlinear-equations, and these preliminary estimates can be used as starting points of the computational methods for maximizing the closed form of the log-likelihood function.

It is worth noting that the ZIG distribution could have been replaced by a hurdle geometric distribution (see Cragg, 1971) with PMF given by:

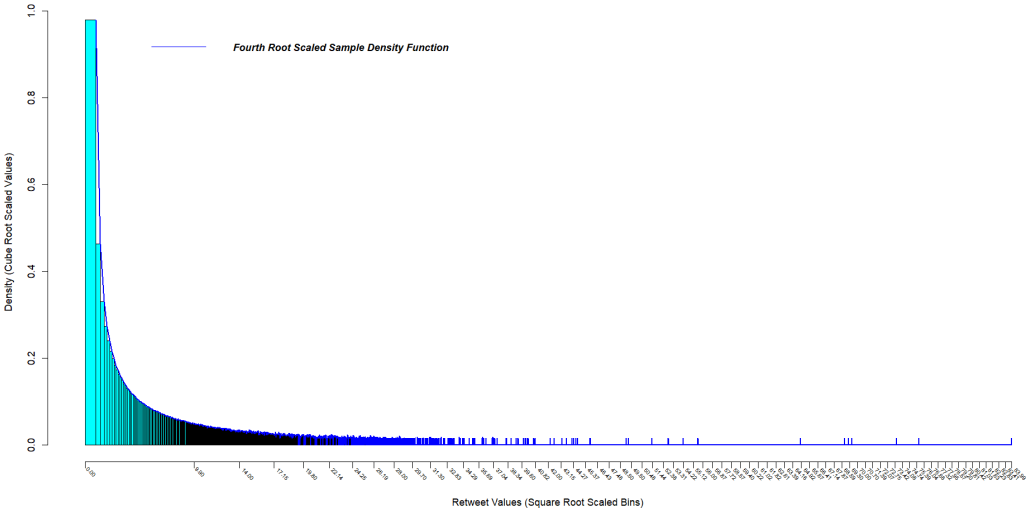
$$f(n|\varphi,p) := \begin{cases} \varphi, & if\ n = 0, \\ (1-\varphi)p(1-p)^{n-1}, & if\ n \in \mathbb{N}, \end{cases} \varphi, p \in (0,1),$$

where φ is the hurdle parameter increasing the probability mass at . In this case, by mixing with the beta distribution we would obtain a hurdle beta geometric (HBG) distribution with PMF given by:

$$f(n|\varphi,\alpha,\beta) := \begin{cases} \varphi, & if\ n = 0, \\ (1-\varphi)\dfrac{\mathrm{B}(\alpha+1,\beta+n-1)}{\mathrm{B}(\alpha,\beta)}, & if\ n \in \mathbb{N}, \end{cases} \varphi \in (0,1), \alpha, \beta > 0,$$

The HBG distribution would offer a simpler structure than the ZIBG distribution, while still providing closed forms for the first three moments. However, for modeling situations with a high number of zero retweets, the ZIBG distribution seems more appropriate to us. This is because the HBG distribution is more suitable for scenarios in which excess zeros stem from a single process, specifically, in our case, some tweets will not be retweeted at all, while all other tweets are guaranteed to be retweeted according to a geometric distribution. In contrast, the ZIBG distribution accommodates scenarios where excess zeros are generated by two processes, in our case, some tweets will never be retweeted, while others may have the potential to be retweeted but do not get retweeted. Considering these key characteristics of the two distributions, we have chosen to use the ZIBG distribution for its superior flexibility.

Figure 1. Density of the Italian retweets from December 1st to December 12, 2020



Fonte: nostre elaborazioni su dati estratti da Twitter

Figure 1 shows the density of retweets in Italian tweets from December 1st to December 12th, 2020.

In addition to modeling the number of retweets, we also consider the virality time, defined as the period during which content spreads rapidly and extensively across a social network or the Internet. This concept encompasses several key aspects: the speed at which the content is shared, the peak moment of its diffusion, and the overall duration of its viral cycle. In our analysis, the unit of measurement for virality time is the hour.

4. Sentiment, multimedia and topics of most rretweeted messages

We analyze the most retweeted tweets, focusing on those with at least 1,000 retweets, representing approximately 1% of all tweets (349,491 tweets). These messages display a strong sentiment bias: 52% carry a negative sentiment, 43% convey positive sentiment, and only 5% are classified as neutral. Additionally, 9% of these tweets include at least one emoji (see Table 1). The sentiment was manually annotated because there were often colloquial expressions or ironic intent, and thus an unsupervised classification based on a dictionary (Liu, 2015) or model-based approach, VADER (Valence Aware Dictionary and sEntiment Reasoner - Hutto and Gilbert, 2014) would not have produced high-quality results. A supervised classification (see Nalini *et al*., 2023) needs a training set, which we can use like Sentiment140 1 to apply machine learning and deep learning algorithms. Still, in those cases, the accuracy of manual classification is superior to that

of automated classification, although it requires a more significant investment of time and human resources. This manual classification has also been used to identify the use of multimedia. We observed that the use of videos and audio is always associated and accounts for 16%, while photos are much more frequent, accompanying 48% of the most retweeted tweets. Table 1 summarizes the sentiment and the use of multimedia tools in the most retweeted tweets.

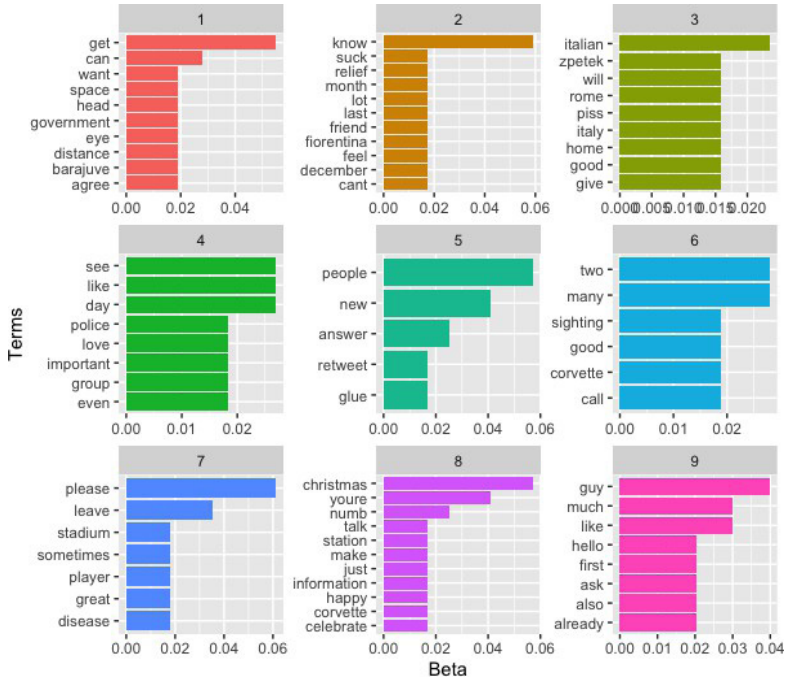
Table 1. Sentiment and multimedial contents in in the most retweeted tweets

Category	Percentage
Positive	43%
Negative	52%
Neutral	5%
Emoje	9%
Photos	48%
Audio + Video	16%

Fonte: nostre elaborazioni su dati estratti da Twitter

To analyze tweet content, we applied the Latent Dirichlet Allocation (LDA) algorithm, a probabilistic model that identifies hidden thematic structures in a text corpus. LDA (see Blei *et al*., 2003) each document (in this case, a tweet) is a mixture of topics, and each topic is a distribution over words. To improve model accuracy, we performed extensive preprocessing, including the removal of URLs, mentions, hashtags, special characters, and numbers, as well as lemmatization. Figure 2 shows the top terms associated with each latent topic extracted through topic modeling (LDA). On the x-axis, Beta represents the estimated probability that a given word belongs to a specific topic. The higher the Beta value, the more representative the word is of that topic.

Figure 2. Top terms for each topic



Fonte: nostre elaborazioni su dati estratti da Twitter

To determine the optimal number of topics, we used perplexity as an evaluation metric: lower perplexity values indicate a better model fit. The final model identified nine topics, each characterized by the most representative terms:

1. Football – tweets centered on football, with frequent terms like "Barajuve", referring to FC Barcelona and Juventus fans.
2. COVID-19 Holiday Restrictions – messages reflecting sadness and isolation due to containment measures during the Christmas period.
3. Lockdown Behavior in Italy – tweets about rediscovered hobbies such as watching Ferzan Özpetek's films during lockdown.
4. Social and Political Engagement – encouragement for solidarity and activism, with expressions like "daleparasempre".
5. Humor and Comedy – lighthearted tweets with jokes and expressions like "hahahaha".
6. Advertising Content – promotional tweets about products, destinations, and cars, such as the Chevrolet Corvette or Braies in South Tyrol.
7. Tributes to Paolo Rossi – condolences and tributes to the legendary footballer.
8. Politics and COVID-19 Policies – criticism of holiday and vaccination policies.
9. Consumer Behavior – tweets about purchases, often online, including platforms like Spotify.

Table 2 reports the main topics identified by the LDA model, together with their thematic interpretation and relative prevalence in the corpus. Most topics account for around 9-11% of the documents, while Topic 9 "Consumer Behavior" is the most represented, covering over one-fifth of the dataset.

Table 2. Topic prevalence and thematic labeling in the Twitter corpus

Topic n°	Description	Dimension (%)
1	*Football*	9.71
2	*COVID-19 Holiday Restrictions*	10.29
3	*Lockdown Behavior in Italy*	8.57
4	*Social and Political Engagement*	9.71
5	*Humor and Comedy*	9.71
6	*Advertising Content*	9.71
7	*Tributes to Paolo Rossi*	10.86
8	*Politics and COVID-19 Policies*	10.29
9	*Consumer Behavior*	21.14

Fonte: nostre elaborazioni su dati estratti da Twitter

Interestingly, the density distribution of followers shows that some highly retweeted tweets originate from accounts with modest followings. However, notable exceptions include Cristiano Ronaldo (110.8M followers), Manchester United (37.7M), and Ibai (13.9M), all linked to the sports sector – highlighting the correlation between virality and sports-related content.

5. Community detection on Twitter/X Heterogeneous Graph

We model user-hashtag interactions using a Heterogeneous Graph $G = \{V, E, \tau, \phi\}$, where V and E are the sets of nodes and edges, respectively, and the functions τ: $V \rightarrow A$ and ϕ: $E \rightarrow R$ map edges in edge types A and nodes in node types R, respectively (retweets RT and hashtag usage M) (Sun *et al*., 2011, 2022). The graph is represented by a symmetric weighted adjacency matrix W with non-negative integer entries, where $W_{i,j} > 0$ if and only if $(i, j) \in E$, and $W_{i,j} = 0$ otherwise. This graph structure captures both relational semantics and content dynamics, enabling a richer representation of social interactions. Retweeting, a key mechanism of information diffusion (Suh *et al*., 2010), reflects the communicative value of content (Cha *et al*., 2010), while hashtag usage supports user visibility (Wang *et al*., 2016) and affiliation with thematic communities (Bruns & Burgess, 2011; Small, 2011; Laucuka, 2018).

To detect communities combining users and hashtags, we tested Louvain (Blondel *et al*., 2008) and Leiden (Traag *et al*., 2019). While both aim to maximize modularity, Leiden was preferred for producing smaller, better-connected clusters and avoiding the oversized partitions typical of Louvain. Other algorithms allowing overlapping communities, such as Conga (Gregory, 2008) and Combo (Sobolevsky *et al*., 2014), were excluded for simplicity.

Using Leiden, we identified over 2 million communities, of which ~90% were singletons. Only 53 had more than 50 nodes. The largest 10 communities (3,795 nodes total) featured distinct thematic areas. For example, communities C_0 and C_9 lacked hashtags (possibly due to the limited time window), while C_1, C_4, and C_6 centered on reality TV (e.g., *Grande Fratello*), and C_3 focused on Turkish TV series. Political content appeared in C_7 and C_8, while COVID-19 and government measures dominated C_2. Community C_5 was more introspective, with users sharing quotes and reflections.

A summary of the top hashtags for the first 10 communities is provided in Table 3.

Table 3. The 10 largest communities by number of nodes, with the most 10 used hashtags by weight

Community	Most 10 used hashtags	# nodes
C_0	N/A	1163
C_1	gfvip, verissimo, secondavita, oppininstudio	676
C_2	mes, conte, covid19, natale, dpcm, m5s, salvini, governo, vaccine	539
C_3	canyaman, özgegürel, mrwrong, produawards2020mrwrong, produawards2020canyaman, canyamanmanoftheyear2020, produawards2020, produawards2020özgegürel, wemissyouözgegürel, bayyanlışrewind	267
C_4	rosmello, dayane, rosalinda, rosmelloilnostrogf, rosmellosempreconvoi	238
C_5	ventaglidiparole, buongiorno, avreivoluto, untemaalgiorno, unsogno, buongiornoatutti, uninvitoa, tuttequellecoseche, cosaèsuccesso, buonanotte	212
C_6	gregorelli, zorzelli, zorpini, pierpaolopretelli, elisabettagregoraci, gregorando, gregorellidellanotte	189
C_7	renzi, report, meloni, lega, reportrai3, fontana, berlusconi, fdi, lanotizia	174
C_8	ottoemezzo, gruber, italiaviva, boschi, philipmorris, casaleggio, cinquestellopoli	169
C_9	N/A	168

Fonte: nostre elaborazioni su dati estratti da Twitter

6. Regression models for over-dispersed count response

To investigate the factors influencing retweets, we applied various regression models (see Cameron and Trivedi, 1990; Korosteleva, 2018, using explanatory variables such as sentiment, the use of multimedia elements like photos, videos, and audio, the number of followers, likes, and network statistics. Retweet data, although they are count data, exhibit overdispersion, meaning that the variance exceeds the mean. For this reason, standard Poisson regression models may not be appropriate. This is because Poisson regression assumes that the mean and variance of the response variable are equal, which is not the case in situations of overdispersion. In this case, we can also use the Zero-Truncated Negative Binomial (ZTNB) regression model, which is used when the response variable is count data that is strictly positive (i.e., there are no zero counts). This model is particularly useful in scenarios where the occurrence of zeros is impossible or does not make sense, such as the number of times an event occurs after it has been triggered. Response Variable: The response variable Y follows a zero-truncated negative binomial distribution. We apply three regression models: Poisson (POIS), Negative Binomial (NB), and ZTNB regression model, using eight distinct models for each type by adding one regressor at a time. Specifically, the regressors encompass two categories of information: content (including audio/video presence, photos, emojis, sentiment, and topics) and network characteristics of the tweeters (followers, hashtags, and clusters).

Table 4. Results Regression models: POIS, NEGB, ZNEGB

Akaike Information (AIC)	POIS	NB	ZTNB
audio/video	137769	1677	1814
audio/video + photo	135383	1676	1813
(audio/video + photo + emoje) = multimedia	121399	1653	1793
multimedia+sentiment	112474	1652	1788
multimedia+sentiment+topic	59769	1633	1790
multimedia+sentiment+topic+follower	59014	1634	1635
multimedia+sentiment+topic+follower+hashtag	59014	**1630**	1631
multimedia+sentiment+topic+follower+hashtag+cluster	59016	1632	1633

Fonte: nostre elaborazioni su dati estratti da Twitter

To identify the most effective statistical models suited to our data, we utilized the Akaike Information Criterion (AIC) for comparison. As demonstrated in Table 4, the NB regression model outperforms all other models when the explanatory variables in-

clude multimedia, sentiment, topic, followers, and hashtags. Additionally, the ZTNB model shows improvement and approaches NB values when network measures are incorporated. These results underscore the benefits of a mixed approach, suggesting that incorporating both content and network variables enhances model performance.

7. Conclusions

This study offers a comprehensive examination of virality on Twitter, showing that the diffusion of messages is shaped by both content features and network structures. The retweet distribution follows a zero-inflated beta-geometric model, which captures the heterogeneity of tweet performance. Measuring virality time in hours provides a more accurate description of the life cycle of viral messages, reflecting the rhythm of their acceleration, peak, and decline.

Content analysis reveals a slight dominance of negative sentiment among the most retweeted tweets, while positive messages remain highly represented and neutral tones are marginal. Multimedia plays a significant role, with photos present in nearly half of the most viral tweets and videos or audio reinforcing engagement despite their lower frequency. Topic modeling highlights a heterogeneous set of themes, with consumer behavior, sports, politics, and COVID-19 emerging as the most influential areas of discussion. Notably, virality does not depend exclusively on large audiences: while sports celebrities and organizations dominate the extremes, smaller accounts can also generate substantial diffusion when content resonates strongly.

Community detection shows that online conversations are organized into a limited number of large thematic clusters, primarily around politics, entertainment, and the pandemic. Finally, regression analyzes confirm that standard Poisson models are inadequate, while Negative Binomial and Zero-Truncated Negative Binomial models perform better, especially when combining content and network predictors. This mixed approach consistently improves explanatory power and predictive accuracy.

In conclusion, virality emerges as the outcome of complex interactions between message design, network configuration, and temporal dynamics. Our findings underline the need for integrated models that account for these dimensions simultaneously, providing insights for both theory and practice. Future research should extend the analysis to multiple platforms, investigate the amplifying role of algorithms, and refine sentiment detection methods to better capture nuances such as irony and sarcasm.

Bibliografia

Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. *Nature, 628*(8008), 582-589. https://doi.org/10.1038/s41586-024-07229-y.

Bene, M. (2017). Go viral on the facebook! Interactions between candidates and followers on Facebook during the Hungarian general election campaign of 2014. *Information, Communication & Society, 20*(4), 513-529.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192-205.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17-35. https://doi.org/10.1214/07-AOAS114.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bruns, A., & Burgess, J. (2011). The use of Twitter hashtags in the formation of ad hoc publics. *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, 1-9.

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*(3), 347-364.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in Twitter: The million follower fallacy. *Proceedings of the International AAAI Conference on Web and Social Media, 4*(1), 10-17.

Chou, P., Chuang, H. H.-C., Chou, Y.-C., & Liang, T.-P. (2022). Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. *European Journal of Operational Research, 296*(2), 635-651.

Cragg, J.G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica, 39(5), 829-844,* https://doi.org/10.2307/1909582.

Dimmock, N., Easton, A., & Leppard, K. (2016, January). *Introduction to modern virology* (7th ed.). Blackwell Publishing.

Elmas, T., Stephane, S., & Houssiaux, C. (2023). Measuring and detecting virality on social media: The case of Twitter's viral tweets topic. *Companion Proceedings of the ACM Web Conference 2023*. https://doi.org/10.1145/3543873.3587373.

Ghaisani, A. P., Handayani, P. W., & Munajat, Q. (2017). Users' motivation in sharing information on social media. *Procedia Computer Science, 124*, 530-535. https://doi.org/10.1016/j.procs.2017.12.186.

Gregory, S. (2008). A fast algorithm to find overlapping communities in networks. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 408-423.

Heimbach, I., & Hinz, O. (2016). The impact of content sentiment and emotionality on content virality. *International Journal of Research in Marketing, 33*(3), 695-701.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 94-100.

Kim, J. W. (2018). They liked and shared: Effects of social media virality metrics on perceptions of message influence and behavioral intentions. *Computers in Human Behavior, 84*, 153-161.

Korosteleva, O. (2018, December). *Advanced regression models with SAS and R*. CRC Press - Taylor Francis Group. https://doi.org/10.1201/9781315169828.

Kulkarni, S., & Rodd, S. F. (2020). Context aware recommendation systems: A review of the state of the art techniques. *Computer Science Review, 37*, 100255. https://doi.org/10.1016/j.cosrev.2020.100255.

Laucuka, A. (2018). Communicative functions of hashtags. *Economics and Culture, 15*(1), 56-62.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789.

Nalini, C., Dharani, B., Baskar, T., & Shanthakumari, R. (2023). Review on sentiment analysis using supervised machine learning techniques. In A. Abraham, S. Pllana, G. Casalino, K. Ma, & A. Bajaj (Eds.), *Intelligent systems design and applications* (pp. 166-177). Springer Nature Switzerland.

Ngo, T. T. A., Bui, C. T., Chau, H. K. L., & Tran, N. P. N. (2024). Electronic word-of-mouth (eWOM) on social networking sites (SNS): Roles of information credibility in shaping online purchase intention. *Heliyon, 10*(11).

Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research, 44*(4), 333-348.

Small, T. A. (2011). What the hashtag? A content analysis of Canadian politics on Twitter. *Information, Communication & Society, 14*(6), 872-895.

Sobolevsky, S., Campari, R., Belyi, A., & Ratti, C. (2014). General optimization technique for high-quality community detection in complex networks. *Physical Review E, 90*(1), 012811.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. *2010 IEEE Second International Conference on Social Computing*, 177-184.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). PathSim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment, 4*(11), 992-1003.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2022). Heterogeneous information networks: The past, the present, and the future. *Proceedings of the VLDB Endowment, 15*(12).

Tiago, F., Moreira, F., & Borges-Tiago, T. (2019). YouTube videos: A destination marketing outlook. *Strategic Innovative Marketing and Tourism: 7th ICSIMAT, Athenian Riviera, Greece, 2018*, 877-884.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 1-12.

Wang, R., Liu, W., & Gao, S. (2016). Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online Information Review, 40*(7), 850-866.

We Are Social & Meltwater. (2024). *Digital 2023 global overview report*. Retrieved from https://datareportal.com/reports/digital-2024-global-overview-report.

DATA SCIENCE E AI: UNA PROPOSTA TECNOLOGICA PER POTENZIARE I PROCESSI DI INNOVAZIONE STATISTICA

DATA SCIENCE AND AI: A TECHNOLOGY PROPOSAL TO IMPROVE STATISTICAL INNOVATION PROCESSES

Francesco Altarocca[1], Domenico Aprile[2], Simonetta Cozzi[3], Armando D'Aniello[4], Annunziata Fiore[5], Enrico Orsini[6], Andrea Pagano[7]

Sommario

In un contesto data-driven in rapida evoluzione, questa ricerca propone come obiettivo una soluzione tecnologica innovativa, basata su strumenti low-code di Data Science e Intelligenza Artificiale (AI), al fine di supportare e ottimizzare i processi di innovazione statistica. L'originalità dello studio risiede nell'impiego della piattaforma low-code RapidMiner che abilita utenti non esperti a costruire, validare e confrontare modelli di Machine Learning (ML) complessi tramite interfacce visuali. L'approccio metodologico prevede un processo strutturato in quattro fasi (acquisizione, preparazione, modellazione e analisi), con particolare attenzione all'automazione delle pipeline, alla costruzione di modelli di Machine Learning mediante diverse metodologie, alla loro ottimizzazione e alla comparazione delle prestazioni.

Il caso d'uso, sperimentato in ISTAT, riguarda l'esplorazione di nuove tecniche di integrazione tra registri statistici per processi di produzione in corso di costruzione ed è finalizzato ad associare le attività economiche delle imprese alle relative unità immobiliari, sfruttando variabili catastali, economiche e territoriali.

La sperimentazione ha evidenziato l'effettiva eliminazione delle tradizionali fasi di implementazione di algoritmi e modelli, nonché l'aumento dell'efficienza nell'allocazione delle risorse computazionali.

Le implicazioni di ricerca indicano che l'adozione di strumenti low-code può democratizzare l'accesso all'AI nelle organizzazioni, supportando l'evoluzione dei processi statistici verso soluzioni più rapide e collaborative, trasparenti e riproducibili.

[1] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: fraltaro@istat.it
[2] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: aprile@istat.it
[3] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: sicozzi@istat.it
[4] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: armando.daniello@istat.it
[5] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: annunziata.fiore@istat.it
[6] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: eorsini@istat.it
[7] Istat, Istituto Nazionale di Statistica, Roma, Italia - e-mail: andrea.pagano@istat.it

Abstract

In a rapidly evolving data-driven context, the aim of this research is to propose an innovative technological solution based on low-code Data Science and Artificial Intelligence (AI) tools in order to support and optimize statistical innovation processes. The originality of the study lies in the use of low-code platforms such as RapidMiner, which enable non-expert users to build, validate, and compare complex machine learning (ML) models through intuitive visual interfaces. The proposed methodological framework consists of a structured four-phase process: data import, data preparation, modeling, and analysis, with a specific focus on pipeline automation, construction of ML models using various methodologies, model optimization, and comparative evaluation.

The use case, tested in the ISTAT context, concerns the exploration of new methods of integration between statistical registers for production processes under construction and aims to associate the economic activities of companies with the relevant real estate units, using cadastral, economic, and territorial variables.

The experiment showed that the traditional phases of implementing algorithms and models were effectively eliminated and that the allocation of computational resources became more efficient.

The research findings suggest that low-code tools have the potential to democratize access to AI within organizations, facilitating the transformation of statistical processes into faster, more collaborative, transparent, and reproducible workflows.

Parole chiave: data science, artificial intelligence, machine learning, low-code.
Keywords: *data science, artificial intelligence, machine learning, low-code.*

1. Introduzione

In una società sempre più caratterizzata da un approccio data-driven, la generazione e l'accumulazione di enormi quantità di dati, strutturati e non, rappresentano per istituzioni, organizzazioni e cittadini un'opportunità per produrre informazioni a supporto delle decisioni strategiche. Parallelamente, l'avanzamento dei software no-code e low-code ha abilitato l'utilizzo di questi dati per fornire approfondimenti, impiegando tecniche e metodologie avanzate di analisi.

Queste piattaforme di analisi dati integrano modelli statistici, metodologie avanzate, tecniche di AI, creazione di pipeline di dati e automazione dei processi, riducendo sensibilmente o eliminando del tutto la necessità di scrivere codice. Ciò permette anche a chi non è particolarmente esperto di implementare facilmente un processo di analisi complesso.

Da questo punto di vista si assiste, da un lato, ad un ampliamento degli strumenti e modelli a disposizione dei ricercatori e, dall'altro, ad una semplificazione, sistematizzazione, standardizzazione e automazione dei processi che porta, di conseguenza, ad una significativa contrazione dei tempi.

Da anni in letteratura è possibile riscontrare diversi contributi e contesti che utilizzano questo tipo di strumenti.

Ad esempio, nel lavoro (Yadav, Malik, & Chandel, 2015) è stato proposto l'uso di RapidMiner per controllare i processi di sperimentazione, per l'import e l'export dei dati, per la data trasformation, per la modellazione e per la valutazione dei risultati. Inoltre, è stato utilizzato per la Principal Component Analysis (PCA), la selezioni delle variabili rilevanti e per il confronto dei diversi modelli di Machine Learning (ML): Artificial Neural Network (ANN), Radial Basis Function Neural Network (RBFNN) e Generalized Regression Neural Network (GRNN).

Nello studio (Vyas, & Uma, 2018), che confronta 20 strumenti per la determinazione del sentiment di tweet, RapidMiner è stato impiegato per determinare la polarità dei tweet utilizzando modelli di classificazione Support Vector Machine (SVM), Decision tree e Naive Bayes. In particolare, nelle conclusioni vengono evidenziate, oltre all'efficienza, la semplicità d'uso e portabilità che lo distinguono dagli altri strumenti utilizzati nello studio.

Come nel caso del presente contributo anche in (Ngadiron *et al*., 2024) sono stati utilizzati e comparati diversi approcci di ML utilizzando RapidMiner. In particolare, nel lavoro sono stati implementati gli algoritmi k-Nearest Neighbors (k-NN), Decision Trees (DT), Deep Learning (DL) e SVM al fine di identificare e creare un modello predittivo per gli incidenti nei cantieri ferroviari.

Anche l'Istat, negli ultimi anni, ha investito notevoli risorse nell'adozione di tecnologie innovative al fine di gestire al meglio il proprio patrimonio informativo. In particolare, ha delineato una roadmap per la costruzione di un ecosistema di piattaforme abilitanti, che integrino anche le più recenti e attuali componenti di AI. In questo percorso, l'Istat si è dotato di strumenti per la prototipizzazione rapida a supporto dei data scientist, oltre ad avanzati strumenti tecnologici innovativi per la Data Virtualization, per la gestione di nuove fonti e big data, per lo sviluppo low-code.

Inoltre, in Istat è presente il laboratorio di Innovazione (LabInn), un ambiente adatto a rafforzare il ruolo della ricerca e a facilitare la nascita e lo sviluppo dell'innovazione, offrendo la possibilità di sperimentare, in modo agile, nuove idee che provengono "dal basso" nell'ambito del programma d'innovazione. Tale programma definisce le aree di priorità considerate innovative e di maggiore interesse per le attività d'Istituto:

- nuove fonti di dati, big data e nuove modalità di acquisizione;
- miglioramento dei processi statistici, adozione di standard e linee guida fornite

da iniziative internazionali;
- nuove tecniche di navigazione, scoperta e visualizzazione dell'informazione, integrazione dati, open data, linked open data;
- nuove tecnologie e metodologie ICT.

Tra i diversi casi d'uso sperimentati con queste tecnologie, quelli di maggiore interesse per grado di innovazione, potenziale evolutivo e capacità di accelerare i processi, riguardano la costruzione di modelli di machine learning supervisionato.

I modelli di machine learning possono essere integrati nel processo di produzione statistica in diversi punti. In particolare, nel caso d'uso analizzato in questo lavoro, la sperimentazione ha permesso di ottimizzare soprattutto le fasi del processo statistico di preparazione, modellazione e integrazione dei dati (fase *Integrate Data*, processo 5.1 del GSBPM[8]), con una significativa riduzione dei tempi complessivi. Le attività che, con un approccio tradizionale fondato sullo sviluppo manuale di codice, avrebbero richiesto diversi mesi sono state completate in poche settimane. Poiché, allo stato attuale dello studio, non sono disponibili baseline comparabili, l'efficacia può essere valutata esclusivamente nei termini qualitativi sopra indicati.

Il Sistema Integrato dei Registri (SIR), introdotto da Istat nel 2016, rappresenta un'infrastruttura statistica centrale per la produzione di statistiche ufficiali. Esso integra registri alimentati da fonti amministrative e indagini campionarie, creando un ambiente informativo unico, strutturato e coerente. Il SIR consente una gestione unitaria e multifunzionale delle informazioni statistiche, garantendo coerenza temporale e tematica tra le diverse fonti. In particolare, si evidenziano il Registro Statistico di Base dei Luoghi (RSBL), che contiene informazioni territoriali sugli edifici, sulle unità immobiliari e sugli indirizzi relativi alle unità statistiche del SIR (unità economiche e individui), e il Registro Statistico delle Imprese Attive (ASIA).

Quest'ultimo raccoglie e aggiorna annualmente i dati relativi alle unità economiche attive (imprese e unità locali) operanti nei settori industriali, commerciali e dei servizi alle imprese e alle famiglie. Esso fornisce informazioni identificative (denominazione e localizzazione) e strutturali (attività economica, numero di addetti dipendenti e indipendenti, forma giuridica, data di inizio e fine attività, fatturato) di tali unità. Ai fini della produzione dell'informazione statistica, le imprese registrate in ASIA sono classificate in base all'attività economica prevalente, definita attraverso la classificazione ATECO. La classificazione ATECO **è** una classificazione gerarchica articolata su sei livelli**,** che vanno dal più generale, contenente ampi raggruppamenti di attività economiche, al più dettagliato, riferito alle singole attività specifiche. Le attività economiche sono orga-

[8] Per ulteriori approfondimenti sul modello Generic Statistical Business Process Model fare riferimento a https://unece.org/sites/default/files/2023-11/GSBPM%20v5_1.pdf

nizzate, dal generale al particolare, secondo la seguente struttura: Sezioni, Divisioni, Gruppi, Classi, Categorie, Sottocategorie**.** Questo sistema codificato consente di classificare in modo uniforme tutte le imprese, favorendo la comparabilità dei dati economici e statistici a livello nazionale ed europeo.

Nell'ambito dell'esplorazione di nuove metodologie di integrazione tra registri statistici per i processi di produzione, e in particolare nel progetto sviluppato all'interno del LabInn, è stata analizzata la possibilità di sviluppare modelli di machine learning basati sulle informazioni catastali, economiche e territoriali disponibili sulle imprese. L'obiettivo è associare le diverse sedi dell'unità economica ai corretti dati catastali, garantendo un adeguato livello di qualità.

2. Gli strumenti low-code per la Data Science: caratteristiche e potenzialità

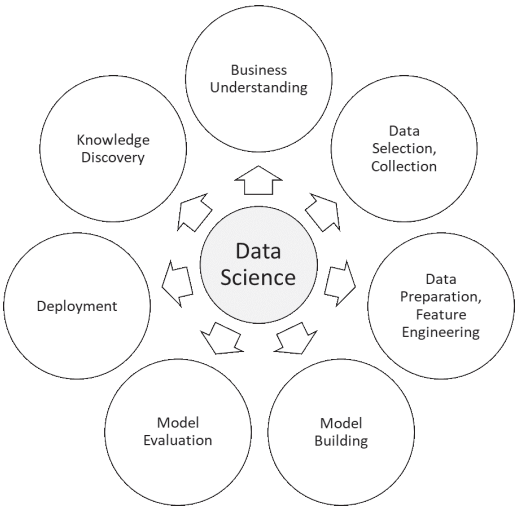
Il ruolo del data scientist nella nostra organizzazione abbraccia molteplici processi e richiede la padronanza di metodologie e tecniche provenienti da diversi ambiti scientifici.

La proliferazione di strumenti, talvolta open source, supportati da una comunità scientifica molto variegata rende questo tipo di attività particolarmente difficile da inglobare in un framework di riferimento coerente e che sia di facile fruizione anche da parte di figure con competenze più orientate al dominio del problema. Documentare tutti gli elementi, le caratteristiche di ciascun oggetto utilizzato e delle strade percorse, è spesso un obiettivo arduo da perseguire.

In questo senso può essere utile l'impiego di strumenti in grado di automatizzare e suggerire percorsi e prove da effettuare, sia per i lavori ripetitivi, come ad esempio la preparazione dei dati, sia per attività con più gradi di libertà come, ad esempio, nell'esplorazione delle scelte delle soluzioni migliori disponibili per un particolare problema.

Molti tool, cui si farà riferimento, sono stati impiegati in processi di produzione, di ricerca e di sperimentazione in Istat, per supportare le operazioni tipiche di questi processi (cfr. Fig. 1). Altri sono promettenti ausili in contesti e casi d'uso che verranno esplorati nel prossimo futuro.

Figura 1. I processi della data science



Fonte: Elaborazione degli Autori

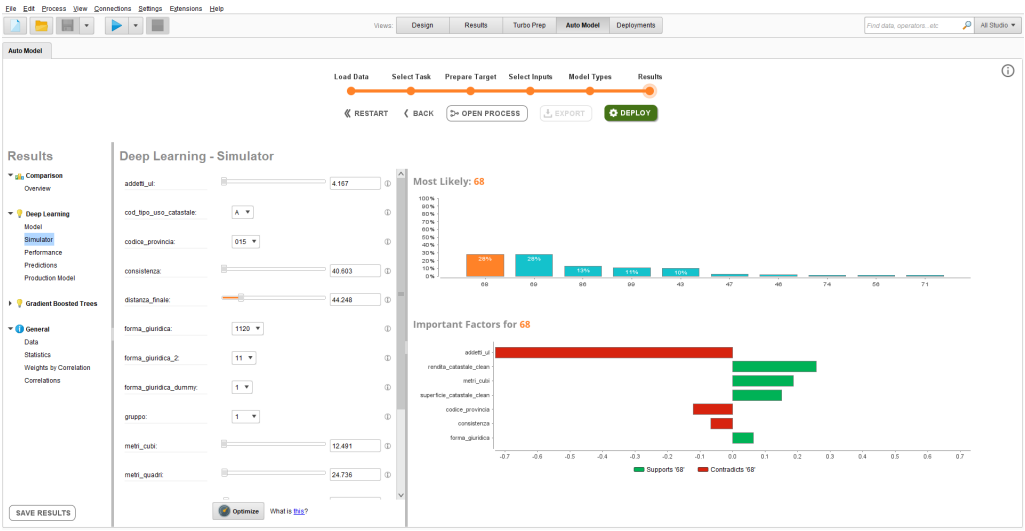
Ad esempio, la preparazione dei dati è una delle fasi che richiede molto tempo. È necessario, infatti, integrare differenti fonti di dati (database, file CSV, Excel, API, cloud) e fare una pulizia dei dati. Tramite componenti che assistono la preparazione dei dati, è possibile automatizzare la gestione dei valori mancanti, la rimozione degli outlier, la trasformazione delle variabili, la riduzione della dimensionalità. Tutto questo è coadiuvato dalla generazione automatica di reportistica di base come, ad esempio, le analisi descrittive delle variabili coinvolte e altre visualizzazioni (grafici, tabelle pivot, heatmap per l'analisi delle correlazioni). Tutte le trasformazioni e le operazioni sono salvate per permettere la riproducibilità e per consentire successivamente la modifica di ciascun blocco della pipeline.

Relativamente alla fase di modellazione e validazione, gli strumenti di modellazione automatica accelerano il processo di costruzione e convalida dei modelli, supportando un'ampia gamma di algoritmi di regressione, classificazione e clustering, tra cui: Random Forest, Support Vector Machines, k-Means, reti neurali, Gradient Boosting Models, modelli di Deep Learning (cfr. Fig. 2) pronti all'uso. Per valutare le performance e compararle tra i diversi algoritmi impiegati, è possibile testare diversi modelli in parallelo ed effettuare il tuning automatico degli iperparametri per capire quali sono le strade più promettenti.

Questi tipi di strumenti sono particolarmente efficaci nell'aiutare il ricercatore, ad esempio, a generare e visualizzare graficamente un processo di grid optimization congiuntamente ai risultati. Consentono inoltre ai team di collaborare, garantendo al tempo

stesso la tracciabilità di tutte le operazioni, fondamentale per la riproducibilità dei processi in fase di deployment: in tal modo contribuiscono ad eliminare operazioni ripetitive, accelerando i tempi di messa in produzione del prototipo, soprattutto nei casi d'uso che prevedono un aggiornamento frequente dei modelli. Questi ed altri vantaggi vengono riportati nelle conclusioni di (Luo, Liang, Wang, Shahin, & Zhan, 2021). Infatti, emerge che strumenti LCD (Low Code Development) forniscono un'interfaccia grafica agevole che consente agli utenti di "programmare" con poco o nessun codice, sono dotati di numerose unità preconfigurate (ad esempio metodi di ML, numerose funzioni statistiche, API e componenti per il trattamento dei dati) facili da imparare e utilizzare, accelerando lo sviluppo e sono particolarmente adatti nei domini che necessitano di processi automatizzati e di flussi di lavoro ripetitivi.

Figura 2. Esempio di funzione di Deep Learning simulator (auto modeling)



Fonte: Piattaforma RapidMiner

3. Caso d'uso: integrazione RSBL con ASIA

L'obiettivo principale della sperimentazione è valutare la possibilità di caratterizzare gli immobili attraverso modelli di machine learning in grado di evidenziare la coerenza tra le categorie catastali e i codici ATECO delle unità locali delle imprese. Tale approccio consente di verificare se esiste un'associazione coerente tra destinazione d'uso degli edifici e attività economica prevalente, al fine di individuare e classificare correttamente gli immobili produttivi rispetto ad altri usi.

La sperimentazione ha riguardato l'utilizzo di modelli di machine learning e di deep learning supervisionati al fine di individuare la migliore collocazione di ciascuna sede dell'unità giuridica presso un immobile preciso sfruttando i dati strutturali dell'impresa (come, ad esempio, ATECO e numero di addetti) e le informazioni catastali dell'immobile (ad esempio superficie, codice catastale e rendita catastale). L'importanza di questa associazione ha impatto sulla qualità e sull'utilizzo dei registri interessati nell'ambito del processo di produzione statistica.

Questa associazione è già determinata attraverso una metodologia deterministica esatta che utilizza i titoli di proprietà o locazione e che verifica la coerenza geografica tra la posizione degli immobili e quella delle unità locali[9] (Istituto Nazionale di Statistica, 2023). La percentuale di associazione relativa al processo deterministico si riferisce a circa il 30% della totalità delle imprese italiane e fornisce un benchmark validato utile per misurare il legame tra le caratteristiche strutturali delle imprese e le informazioni catastali delle unità immobiliari. Utilizzando quindi questo insieme di casi noti, è possibile definire dei modelli ML per imputare un immobile nei casi in cui le singole unità locali siano associate a più di un uno.

Architettura

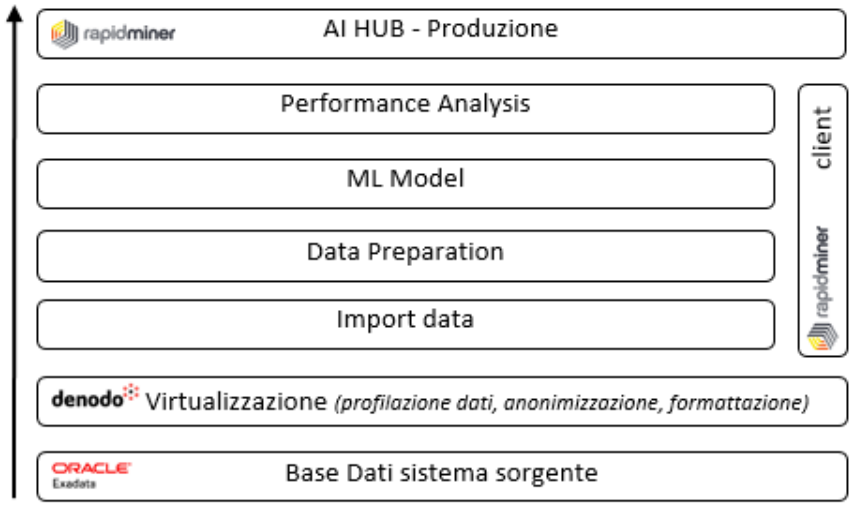
La soluzione proposta si basa su un'architettura (cfr. Fig. 3) integrata che consente di acquisire, trasformare e analizzare i dati in modo sistematico, assicurando la qualità delle informazioni utilizzate per l'addestramento del modello predittivo.

L'architettura implementata si basa sull'integrazione di componenti tecnologiche eterogenee disponibili nel nostro ecosistema tecnologico, configurate per supportare l'intero flusso di elaborazione dati e sviluppo del modello predittivo. Nello specifico la struttura è così composta:

- Oracle: piattaforma database del sistema sorgente;
- Denodo: piattaforma di virtualizzazione dati utilizzata come layer di accesso alle fonti informative necessarie all'analisi;
- RapidMiner: piattaforma di data science utilizzata per l'implementazione dell'intero flusso di data engineering e machine learning e per il successivo deployment ed esecuzione centralizzata dei processi sviluppati.

[9] L'unità locale è il luogo fisico nel quale un'unità giuridico-economica (istituzione) esercita una o più attività economiche. L'unità locale corrisponde ad un'unità giuridico-economica o ad una sua parte, situata in una località topograficamente identificata da un indirizzo e da un numero civico.

Figura 3. Architettura



Fonte: Elaborazione degli Autori

L'impiego di questi strumenti ha reso possibile la realizzazione della soluzione seguendo un approccio strutturato, articolato in quattro macro-fasi interconnesse che coprono l'intero ciclo di vita del caso d'uso. L'intero processo è stato sviluppato tramite interfacce grafiche intuitive, basate su modalità drag & drop, che hanno semplificato significativamente le operazioni. Questo ha permesso, con pochi passaggi, di minimizzare la necessità di conoscenze tecniche approfondite, favorendo l'accessibilità e la rapidità di implementazione.

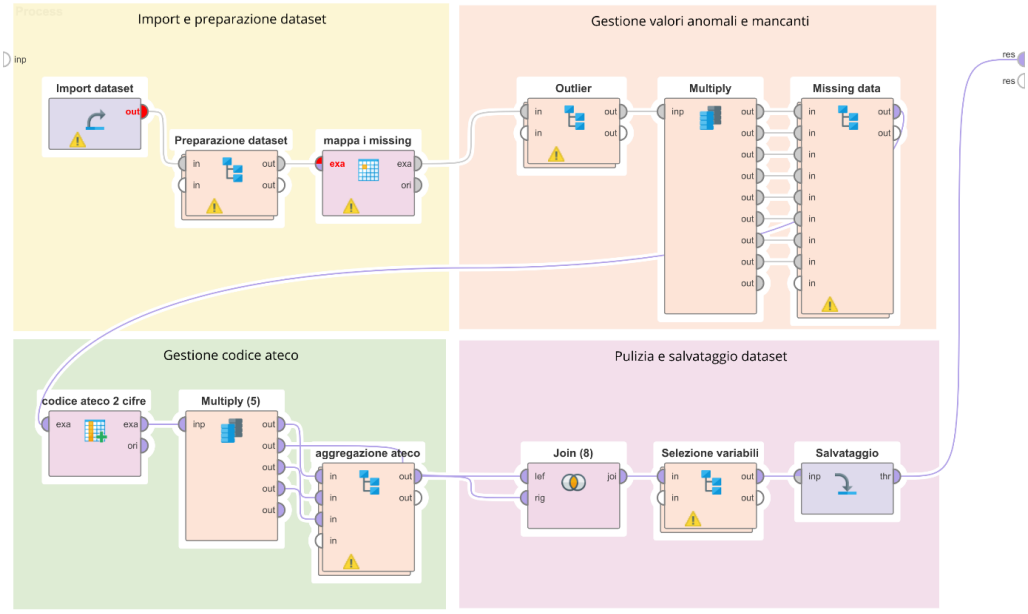
Acquisizione dei dati

In questa prima fase è stata definita la connessione con la piattaforma Denodo per l'estrazione dei dataset di interesse tramite query. La piattaforma Denodo implementa uno strato di virtualizzazione intermedio che mappa, in maniera trasparente, le sorgenti dati memorizzate in questo caso su database Oracle. Uno dei principali vantaggi nell'utilizzo di un connettore unico verso molteplici sorgenti dati è rappresentato dalla possibilità di demandare allo strato di virtualizzazione operazioni fondamentali, come la pseudonimizzazione, la profilazione e altre elaborazioni preliminari, senza dover movimentare fisicamente i dati o introdurre ridondanze. Questo approccio garantisce maggiore efficienza, sicurezza e coerenza nell'accesso e nella gestione delle informazioni.

Preparazione dei dati

Sono state definite delle procedure sistematiche di data preprocessing (cfr. Fig. 4) finalizzate all'ottimizzazione della qualità del dataset.

Figura 4. Esempio di pipeline per la data preparation



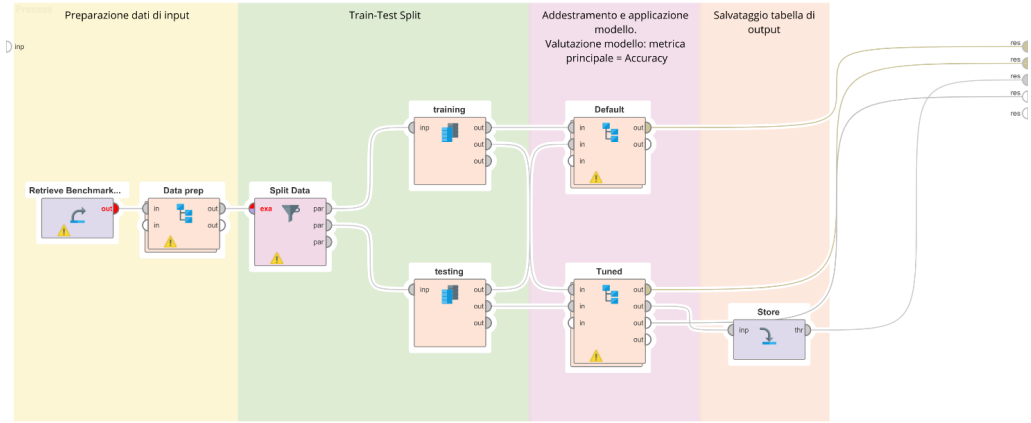
Fonte: Piattaforma RapidMiner

Questa fase ha incluso i seguenti passi: l'identificazione e gestione dei valori anomali, l'implementazione di strategie di imputazione per i valori mancanti, differenziate in base alla tipologia di variabile, la standardizzazione e aggregazione dei codici ATECO alla seconda cifra per ridurne la numerosità e infine la selezione delle variabili più significative per l'addestramento del modello predittivo.

Sviluppo del modello

È stato implementato e addestrato un modello predittivo (cfr. Fig. 5) basato su algoritmi di Gradient Boosted Trees, dopo aver valutato altre tipologie come ad esempio DL.

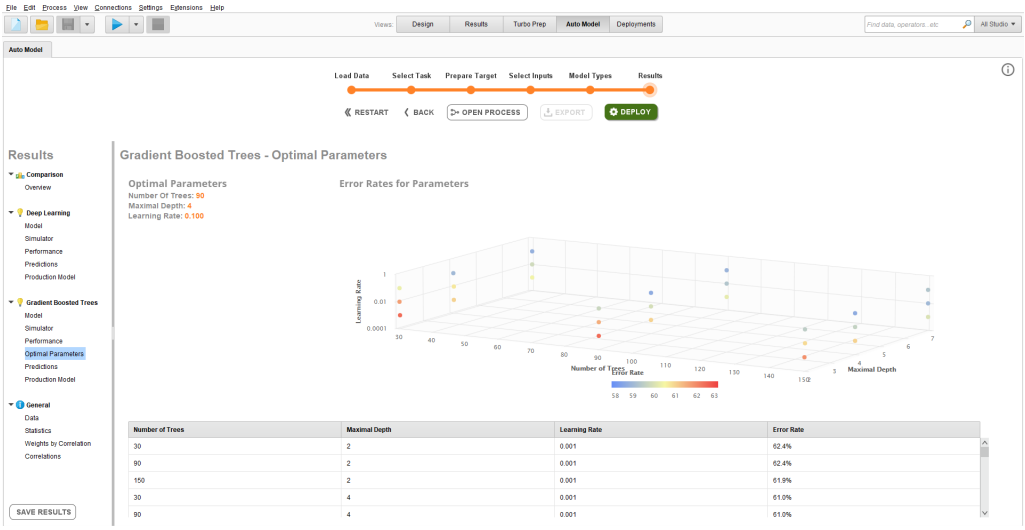
Figura 5. Esempio di pipeline del modello



Fonte: Piattaforma RapidMiner

In particolare, i dati a disposizione sono stati partizionati in training set (80%) e test set (20%). In seguito, oltre alla configurazione dei parametri di ottimizzazione dell'algoritmo (cfr. Fig. 6), sono state implementate procedure di validazione per la valutazione della bontà del modello. Come riportato in figura sottostante, la ricerca dei parametri ottimali del modello è effettuata sulla base di range di valori da assegnare ai parametri: Number of trees, Maximal depth e Learning rate. A ciascun modello è associato un valore di Error Rates che caratterizza la bontà del modello: nella figura i punti colorati da blu scuro a rosso consentono visivamente di visualizzare i risultati ottenuti dalla generazione dei 27 modelli generati (3 Number of trees, 3 Maximal depth e 3 Learning rate diversi).

Figura 6. Esempio di ottimizzazione di Gradient Boosted Trees



Fonte: Piattaforma RapidMiner

Analisi delle performance

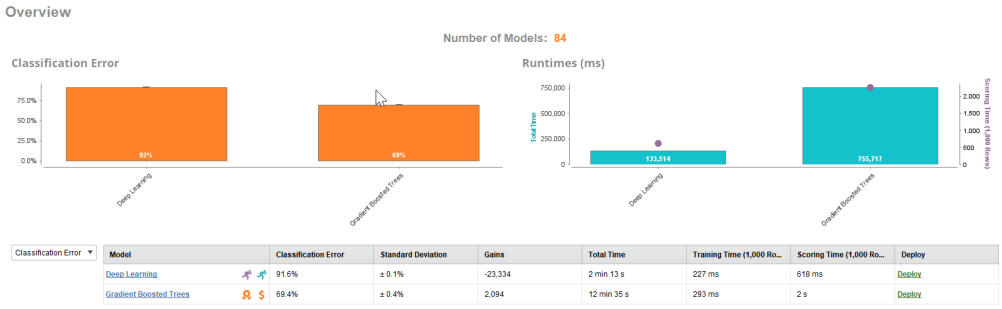
È stata infine effettuata una valutazione approfondita delle metriche di performance del modello, con focus sulla capacità di includere il codice ATECO reale tra le prime tre previsioni con probabilità più elevata. Infatti, il metodo fornisce anche una probabilità di appartenenza ad un codice ATECO. Attraverso altre pipeline di trasformazione è stato creato quindi un dataset contenente le prime 3 associazioni con livello di probabilità maggiore.

Il modello selezionato è un Gradient Boosted Trees, risultato il più performante nelle analisi preliminari condotte attraverso lo strumento AutoModel (cfr. Fig. 7), superando in termini di efficacia anche architetture di tipo neurale.

AutoModel è un'estensione di RapidMiner che accelera il processo di creazione e convalida dei modelli. Consiste nella creazione di un processo che è possibile modificare o mettere in produzione in autonomia nel contesto di tre categorie di problemi: classificazione o regressione, clustering e identificazione di valori anomali. AutoModel, una volta completati i calcoli, guida l'utente nella valutazione e nel confronto dei risultati di differenti modelli[10].

[10] Per ulteriori informazioni fare riferimento alla documentazione ufficiale https://docs.rapidminer.com/2025.1/studio/guided/auto-model/index.html

Figura 7. Comparazione tra modelli



Fonte: Piattaforma RapidMiner

I modelli generati hanno mostrato performance predittive incoraggianti, soprattutto nei casi in cui si estraggono le coppie o le triple con maggiore probabilità. La metrica principale adottata per la valutazione delle performance è l'accuracy, rispetto alla quale sono stati riscontrati i seguenti risultati:

il modello preconfigurato fornito dallo strumento RapidMiner ha registrato un'accuratezza del 41% (errore di classificazione pari al 59%), mentre il modello sottoposto a un processo di fine-tuning ha raggiunto un'accuratezza del 49% (errore di classificazione pari al 51%).

Naturalmente, nell'analisi dei risultati occorre considerare l'elevato numero di classi possibili, corrispondenti a tutti i possibili codici Ateco individuati.

Tali risultati, pur evidenziando margini di miglioramento, confermano la potenziale utilità dell'impiego di modelli di machine learning nel contesto applicativo analizzato, offrendo spunti concreti per l'automazione e l'ottimizzazione dei processi decisionali.

L'ultimo processo ha lo scopo di approfondire l'analisi delle prestazioni del modello, verificando se il codice ATECO reale è incluso tra le prime due predette con la probabilità più elevata.

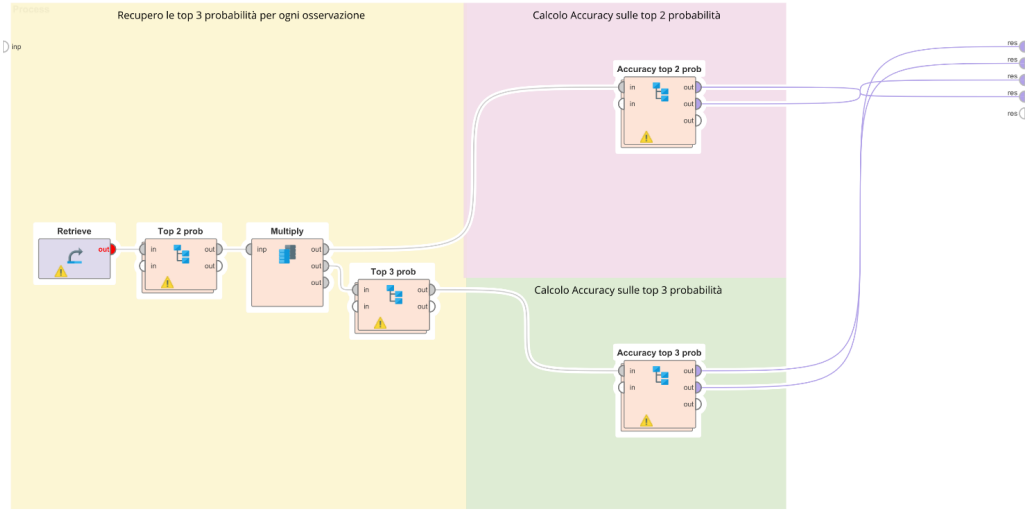
In questa fase del processo vengono considerate, per ciascuna osservazione, le prime due classi con la probabilità più elevata restituite dal modello.

L'obiettivo è valutare se il target corretto sia compreso tra le due alternative più probabili, ampliando così la nozione di accuratezza oltre la sola predizione principale.

Utilizzando il modello ottimizzato tramite fine-tuning, si ottiene in questo scenario una Top-2 accuracy del 71%, corrispondente a un errore di classificazione del 29% (cfr. Fig. 8). Estendendo ulteriormente l'analisi, vengono considerate le prime tre classi a più alta probabilità per ciascuna osservazione. Questa modalità consente di verificare se il valore corretto sia tra le tre ipotesi più plausibili individuate dal modello, offrendo una

misura più flessibile della sua capacità predittiva. Anche in questo caso, utilizzando il modello con parametri ottimizzati, si ottiene una Top-3 accuracy pari all'84%, equivalente a un errore di classificazione del 16%.

Figura 8. Esempio di pipeline dell'analisi delle performance



Fonte: Piattaforma RapidMiner

Sebbene i modelli testati non siano ancora sufficientemente maturi per un utilizzo in produzione, e considerando che è stato analizzato solo un campione, l'output del modello che genera le Top-3 associazioni risulta utile per ridurre significativamente la dimensionalità del problema originario. In particolare, nella matrice di associazione tra il codice ATECO delle unità locali e la categoria catastale dell'immobile, una delle due dimensioni viene ridotta a una cardinalità massima di 3.

4. Generalizzabilità e riproducibiltà della soluzione

Da un lato, dal punto di vista del processo di integrazione dei registri, questa sperimentazione ha permesso di valutare rapidamente numerose tecniche di machine learning, riducendo così i diversi percorsi possibili nel processo di integrazione. Dall'altro, il metodo risulta replicabile nella fase di integrazione dei dati del processo GSBPM ogni volta che si vogliano associare due unità statistiche tramite le loro variabili, come nel caso dell'associazione tra il numero di componenti della famiglia e la superficie dell'alloggio. In generale, la metodologia è riutilizzabile in tutti quei contesti in cui occorre integrare registri o archivi diversi, purché siano presenti variabili di classificazione o indicatori confrontabili (ad esempio codici ATECO, categorie catastali o codici territoriali).

Il processo sviluppato può essere quindi adattato ad altri domini statistici caratterizzati da problemi simili di collegamento e coerenza. Questo rende l'approccio trasferibile non solo all'ambito delle imprese, ma anche ad altri settori della statistica ufficiale. Oltre al conteso appena citato, i modelli di machine learning, e quindi il processo appena proposto, possono essere utilizzati in ampi settori del processo di produzione statistico come, ad esempio, nel controllo e correzione per imputare i valori mancanti e nelle attività di anomaly detection. A supporto di ciò, in (United Nations Economic Commission for Europe, 2022) sono riportati numerosi casi d'uso di tali tecniche nel capitolo "Machine Learning Application Areas". Il crescente interesse verso queste metodologie è ulteriormente confermato da iniziative recenti, tra cui l'intervento presentato nel 2024 nel contesto della Conference of European Statisticians promossa dalla United Nations Economic Commission for Europe (UNECE) (Piela, 2024).

La valutazione quantitativa dell'efficacia dell'approccio low-code presenta alcune peculiarità che meritano di essere esplicitate. Trattandosi di una sperimentazione metodologica innovativa nel contesto ISTAT, non esistono precedenti progetti in questo ambito direttamente comparabili che possano fungere da baseline per un confronto sistematico. Questa assenza di benchmark interni rende difficile la definizione di indicatori di efficacia basati su metriche comparative. Tuttavia, è possibile fornire alcune considerazioni qualitative significative sull'efficienza del processo. Dal punto di vista dell'efficienza temporale, il team interdisciplinare, composto da data scientist, esperti di dominio e tecnici ICT, è riuscito a sviluppare, testare e validare molteplici modelli di machine learning in un arco temporale di qualche settimana. In un approccio tradizionale basato esclusivamente su programmazione, la stessa attività avrebbe richiesto un orizzonte temporale decisamente superiore, dell'ordine di qualche mese, dovuto principalmente a: sviluppo manuale delle pipeline di data preparation, implementazione e test degli algoritmi di machine learning e ottimizzazione, creazione di script personalizzati per la comparazione dei modelli, produzione di documentazione di tutte le fasi e presentazione dei risultati. Questa riduzione dei tempi è stata ottenuta principalmente grazie all'automazione delle fasi ripetitive e alla disponibilità di componenti preconfigurati.

Dal punto di vista dei costi, è importante sottolineare che l'acquisizione della piattaforma RapidMiner è stata effettuata nell'ambito di un investimento strategico più ampio dell'Istituto, finalizzato a supportare più iniziative di innovazione statistica. I costi delle licenze e dell'infrastruttura computazionale sono quindi ammortizzati su diversi progetti e casi d'uso, rendendo l'investimento sostenibile nel medio lungo periodo. Analogamente, le competenze acquisite dal personale attraverso la formazione sulla piattaforma rappresentano un asset riutilizzabile per future sperimentazioni che, nel contesto AI, si stanno recentemente moltiplicando.

Un altro aspetto rilevante emerso dalla sperimentazione riguarda la capacità della piattaforma di facilitare la collaborazione tra figure professionali con competenze diverse. Gli esperti di dominio statistico hanno potuto partecipare attivamente alle fasi di modellazione e validazione, grazie all'interfaccia visuale, senza necessitare di competenze avanzate di programmazione. Questo ha ridotto significativamente i cicli di iterazione e feedback tra team tecnico e analisti statistici.

In sintesi, è opportuno riconoscere che, in assenza di dati comparativi puntuali, le considerazioni qui presentate hanno carattere prevalentemente qualitativo. Studi futuri potrebbero beneficiare di un disegno sperimentale che preveda il confronto sistematico tra approcci low-code e tradizionali su casi d'uso analoghi, al fine di quantificare con maggiore precisione i benefici in termini di tempo, costi e qualità dei risultati.

5. Conclusioni

L'esperienza condotta ha evidenziato risultati significativi sotto molteplici aspetti, confermando il valore aggiunto dell'adozione di piattaforme low-code nell'ambito dell'analisi avanzata dei dati.

Il fine dello use case non era peraltro il raggiungimento del miglior risultato possibile, ma valutare il supporto di una tale piattaforma integrata, in tutte le fasi di realizzazione di una soluzione analitica avanzata.

Si è registrata una notevole contrazione dei tempi necessari per l'implementazione del processo innovativo. L'utilizzo di strumenti basati su interfacce visuali e componenti modulari ha consentito di sviluppare, testare e affinare i modelli con un'efficienza notevolmente superiore rispetto agli approcci tradizionali basati esclusivamente sulla programmazione. Nonostante non siano disponibili metriche puntuali, possiamo affermare che si tratta di una soluzione integrata che consente l'automazione della creazione di modelli, con poca o nessuna scrittura di codice.

Le piattaforme impiegate hanno permesso di effettuare, con estrema semplicità, prove comparative su numerose tecniche, consentendo un'analisi approfondita delle diverse soluzioni, anche grazie alla loro rappresentazione immediata e intuitiva tramite visualizzazioni grafiche integrate nella piattaforma stessa. La semplificazione dei dettagli tecnici più complessi ha permesso al team di concentrarsi sugli aspetti sostanziali dell'analisi statistica e della validazione dei risultati, piuttosto che sulla gestione delle complessità implementative.

Un ulteriore aspetto rilevante dell'esperienza è la possibilità di schedulare in modo efficiente la potenza di calcolo necessaria per l'addestramento dei modelli. Questa funzionalità ha assunto un'importanza cruciale nelle tecniche avanzate di AI, soprattutto di Deep Learning, notoriamente esigenti in termini di risorse computazionali. Infine, la

piattaforma ha consentito un'allocazione ottimale della potenza di calcolo disponibile e ha permesso di generare un gran numero di modelli di AI.

Il processo progettato può essere strutturato come un template riutilizzabile per affrontare progetti analoghi in ambito statistico. Da un punto di vista tecnico, ciò significa che, per applicarlo a una nuova istanza di problema, è possibile mantenere invariato il framework di addestramento e validazione relativo alla batteria di modelli da valutare, mantenendo il connettore e sostituendo le fonti dati ed eventualmente i parametri di configurazione. In tal modo, la soluzione diventa replicabile e adattabile a contesti differenti con interventi minimi sull'architettura complessiva.

I risultati ottenuti attraverso la nostra esperienza trovano ampia conferma nei contributi scientifici presenti in letteratura. In (Ajimati, Carroll, & Maher, 2025), ad esempio, pur in un contesto non del tutto sovrapponibile a quello analizzato in questo contributo, emerge come le organizzazioni sfruttino le piattaforme low-code per ottimizzare i processi aziendali, migliorando l'efficienza in termini di tempo, sforzo, costi e collaborazione. Tali strumenti contribuiscono inoltre a rendere i ricercatori più autonomi e innovativi, aumentando la loro motivazione a partecipare attivamente alla trasformazione digitale delle organizzazioni.

Sebbene l'adozione di questi strumenti abbia portato a numerosi vantaggi, è importante sottolineare che, al momento, risulta ancora difficile sviluppare strategie efficaci per sfruttare appieno le potenzialità delle tecnologie proposte. In particolare, la difficoltà nella definizione di indicatori affidabili e la necessità di significativi investimenti in infrastrutture, competenze professionali e formazione rappresentano una sfida soprattutto nelle fasi iniziali dell'implementazione delle tecnologie low-code nei contesti organizzativi.

Dal punto di vista applicativo, anche alla luce dei risultati promettenti emersi nella fase di sperimentazione, questo caso d'uso apre a numerosi scenari evolutivi e opportunità di implementazione nei processi statistici. Ciò contribuisce ad arricchire e innovare le metodologie di integrazione tra i registri, rafforzando l'efficacia complessiva dei sistemi.

Infine, l'adozione di questi strumenti ha consentito di ampliare in modo significativo il numero di modelli testati su un caso d'uso specifico, senza richiedere un incremento delle risorse disponibili. La convergenza tra i risultati empirici e le evidenze scientifiche documentate rafforza ulteriormente la validità dell'approccio metodologico adottato, inserendolo in un framework teorico consolidato e consentendone l'estensione a contesti di ricerca e produzione in modo sistematico.

Bibliografia

Ajimati, M. O., Carroll, N., & Maher, M. (2025). Adoption of low-code and no-code development: A systematic literature review and future research agenda. *The Journal of Systems and Software*, *222*, Article 112300. https://doi.org/10.1016/j.jss.2024.112300.

Istituto Nazionale di Statistica (2023). *Censimenti permanenti istituzioni pubbliche: Istruzioni per l'individuazione delle unità locali*. Retrieved from https://www.istat.it/wp-content/uploads/2024/02/Istruzioni-per-individuare-Unita-locali_2023.pdf

Luo, Y., Liang, P., Wang, C., Shahin, M., & Zhan, J. (2021). Characteristics and Challenges of Low-Code Development: The Practitioners' Perspective. *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1-11. https://doi.org/10.1145/3475716.3475782.

Ngadiron, Z., Ganasan, R., Ramli, M. F., Mahyeddin, M. E., Luqman, M. I., Jiafu, G., & Kamaluddin, N. A. (2024). Predictive Model for Incident Severity at Railway Construction Site Using Rapid Miner. *International Journal of Sustainable Construction Engineering and Technology*, *15*(4). https://doi.org/10.30880/ijscet.2025.15.04.005.

Piela, R. (2024). Incorporating AI into statistical standards: Enhancing GSBPM with (generative) AI. Presentazione al *ModernStats World Workshop, Commissione Economica per l'Europa delle Nazioni Unite (UNECE)*, Ginevra, Svizzera. Retrieved from https://unece.org/sites/default/files/2024-06/MWW2024_S6_Finland_Piela_A.pdf

United Nations Economic Commission for Europe (2022). *Machine learning for official statistics*. Retrieved from https://unece.org/statistics/publications/machine-learning-official-statistics.

Vyas, V., & Uma, V. (2018). An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner. *Procedia Computer Science*, *125*, 329-335. https://doi.org/10.1016/j.procs.2017.12.044.

Yadav, A. K., Malik, H., & Chandel, S. S. (2015). Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India. *Renewable & Sustainable Energy Reviews*, *52*, 1093-1106. https://doi.org/10.1016/j.rser.2015.07.156.

COME L'ADOZIONE DI UN'IA RESPONSABILE STA CAMBIANDO LE STATISTICHE UFFICIALI

RESPONSIBLE AI ADOPTION: HOW IT'S CHANGING OFFICIAL STATISTICS

Gerarda Grippo[1], Alessandra Righi[2]

Sommario

L'articolo analizza l'impatto dell'intelligenza artificiale (IA) e del machine learning (ML) sull'evoluzione della statistica ufficiale, evidenziandone il potenziale nel migliorare accuratezza, tempestività e profondità analitica dei dati. Queste tecnologie, oltre ad automatizzare processi molto articolati, offrono nuove opportunità per l'analisi di fenomeni complessi e per la costruzione di indicatori innovativi, a supporto delle decisioni pubbliche. La responsabilità nell'uso dell'IA consiste nell'integrare tecnologie avanzate nei processi statistici in modo etico, trasparente e conforme alle normative, garantendo la qualità dei dati, la protezione della privacy e la fiducia del pubblico. Ciò richiede un approccio interdisciplinare che integri competenze tecniche, statistiche, giuridiche ed organizzative, per garantirne un utilizzo responsabile e trasparente. Attraverso l'analisi di esperienze maturate in ambito nazionale e internazionale, il contributo discute le principali sfide metodologiche, organizzative e istituzionali connesse all'introduzione dell'IA nei processi statistici. I primi risultati sperimentali mostrano come l'impiego di queste tecniche, se adeguatamente governato, possa effettivamente contribuire al miglioramento della qualità informativa, anche affrontando criticità come la rappresentatività e i bias nei dati. Il lavoro sottolinea infine l'importanza di una solida governance, di infrastrutture digitali adeguate e dello sviluppo di competenze multidisciplinari, per accompagnare in modo sostenibile la transizione verso una statistica ufficiale potenziata dall'intelligenza artificiale.

Abstract

The article examines the impact of artificial intelligence (AI) and machine learning (ML) on the evolution of official statistics, highlighting their potential to enhance data accuracy, timeliness, and analytical depth. Beyond automating complex processes, these technologies offer new opportunities for analyzing multifaceted phenomena and developing innovative indicators to support public decision-making. Responsibility in the use of AI involves integrating advanced technologies into sta-

[1] Istat - Istituto nazionale di statistica, Roma, Italia - e-mail: grippo@istat.it

[2] Istat - Istituto nazionale di statistica, Roma, Italia - e-mail: righi@istat.it

tistical processes in an ethical, transparent, and legally compliant manner, ensuring data quality, privacy protection, and public trust. This requires an interdisciplinary approach that integrates technical, statistical, legal, and organizational expertise to ensure responsible and transparent use. Drawing on national and international experiences, the paper discusses the main methodological, organizational, and institutional challenges associated with the integration of AI into statistical processes. Initial experimental results show that, when appropriately governed, these techniques can significantly contribute to improving information quality, while addressing critical issues such as data representativeness and algorithmic bias. The paper concludes by emphasizing the importance of robust governance, adequate digital infrastructure, and the development of multidisciplinary skills to support a sustainable transition toward AI-enhanced official statistics.

Parole chiave: innovazione statistica, qualità e governance dei dati, responsabilità algoritmica.
Keywords: *statistical innovation, data quality and governance, algorithmic accountability.*

1. Introduzione

L'applicazione dell'intelligenza artificiale (IA) e del machine learning (ML) negli Istituti nazionali di statistica si sta progressivamente affermando come leva strategica per innovare l'intero ciclo di vita dei dati, dalla raccolta all'elaborazione, fino all'analisi e alla diffusione. Queste tecnologie offrono strumenti potenti per affrontare nuove sfide metodologiche legate all'aumento della complessità informativa e alla disponibilità di fonti eterogenee.

L'impiego di modelli predittivi e classificatori consente di automatizzare attività ad alto dispendio di risorse, come la codifica di testi, la rilevazione di outlier o l'analisi di immagini. Tuttavia, l'integrazione efficace di tali soluzioni non può essere solo tecnologica, richiede infatti un ripensamento complessivo dei processi, delle competenze e delle garanzie istituzionali che regolano la produzione statistica.

Considerando l'impatto che le decisioni basate su modelli possono avere su differenti gruppi sociali e contesti territoriali, è essenziale che i sistemi intelligenti siano regolati da principi quali la responsabilità, l'equità e l'inclusione. Un'adozione consapevole dell'IA in questo contesto implica la definizione di criteri di trasparenza, verificabilità e controllo umano.

Un'ulteriore riflessione va rivolta ai modelli linguistici di grandi dimensioni (Large Language Models, LLM), sempre più utilizzati anche nelle pubbliche amministrazioni e nella produzione statistica sperimentale. Questi modelli – basati su architetture neu-

rali addestrate su enormi corpora testuali – non comprendono realmente il contenuto, ma generano testi probabilisticamente plausibili sulla base di pattern linguistici appresi. Quindi, risposte errate, inventate o fuorvianti generate con apparente sicurezza ma spesso prive di fondamento nei dati reali o nelle fonti disponibili, le cosiddette "allucinazioni", non costituiscono un malfunzionamento, bensì rappresentano una proprietà strutturale del funzionamento statistico di questi sistemi (Loru *et al*., 2025).

Di conseguenza, l'uso di LLM in ambiti in cui è necessario un elevato grado di affidabilità informativa non può prescindere da un solido presidio umano e istituzionale, sia nella validazione dei contenuti, sia nella progettazione dei contesti d'uso. In assenza di tali garanzie, si rischia di introdurre distorsioni e automatismi ingiustificati nella catena di produzione statistica, compromettendo la qualità e l'affidabilità dei risultati.

Ovviamente, la qualità e la solidità dell'infrastruttura informativa su cui poggia l'uso dell'IA riveste un ruolo decisivo. L'adozione di modelli intelligenti richiede architetture flessibili, interoperabili e capaci di garantire la sicurezza, la continuità e la documentazione dei processi. Solo così sarà possibile passare da esperimenti isolati a pratiche strutturali che rafforzino il ruolo pubblico della statistica nel governo dell'informazione. Altrettanto fondamentale per assicurare una gestione etica e sostenibile dei progetti basati su IA è la capacità delle organizzazioni di sviluppare competenze trasversali, costruendo équipe multidisciplinari in grado di coniugare saperi informatici, statistici, giuridici e organizzativi.

In questo lavoro, si esamina il percorso dell'adozione delle nuove tecniche negli Istituti di statistica, dalle fasi sperimentali all'implementazione in produzione. Si discute dei primi risultati dell'adozione di queste tecniche e di come un uso responsabile e trasparente all'IA richieda la promozione di nuove strutture di governance e programmi di sviluppo delle competenze.

2. L'adozione responsabile dell'intelligenza artificiale nella produzione statistica

L'intelligenza artificiale sta trasformando profondamente la produzione statistica, migliorandone l'efficienza operativa attraverso l'automazione dei compiti ripetitivi. Questo consente di alleggerire il carico di lavoro manuale, liberando risorse per attività strategiche e analisi complesse, con un impatto diretto sulla produttività e sulla capacità interpretativa. Inoltre, l'IA amplia le possibilità di integrazione delle fonti, permettendo di combinare dati tradizionali con flussi informativi emergenti, come quelli provenienti dai social media, dai sensori o da piattaforme digitali. Ne derivano dataset più ricchi e una comprensione più profonda dei fenomeni osservati.

Negli ultimi anni, il settore privato ha saputo cogliere con rapidità le potenzialità offerte dall'IA e dal machine learning, sviluppando strumenti statistici accessibili, personalizzati e veloci. Questo dinamismo ha suscitato crescente attenzione da parte di

decisori politici e stakeholder, ponendo alle istituzioni pubbliche la sfida di modernizzarsi sotto il profilo tecnologico e organizzativo. Istituti nazionali di statistica e autorità pubbliche sono così chiamati a rafforzare la propria capacità di innovazione, senza mai compromettere l'integrità metodologica e i principi che contraddistinguono la statistica ufficiale.

L'adozione dell'IA, tuttavia, solleva interrogativi etici rilevanti. L'automazione non può sostituire il giudizio umano, che rimane cruciale soprattutto nelle decisioni sensibili. È quindi necessario definire linee guida per un uso responsabile dell'IA, istituire comitati etici e garantire costantemente la possibilità di supervisione e intervento umano.

La trasparenza algoritmica rappresenta una sfida cruciale, poiché i modelli complessi possono operare come "scatole nere", rendendo difficile comprendere i processi decisionali. Questo fenomeno è strettamente legato al bias algoritmico, che si verifica quando i dati utilizzati per addestrare i sistemi intelligenti sono distorti, generando risposte fuorvianti. Per mitigare tali rischi, è necessario adottare modelli interpretabili, corredati da una documentazione esaustiva e conformi agli standard regolatori, come quelli promossi dall'Unione Europea. Inoltre, è fondamentale ridurre i bias nei sistemi intelligenti attraverso tecniche di de-biasing, garantendo la qualità e la rappresentatività dei dati e monitorando costantemente i risultati.

Tuttavia, oltre alle problematiche tecniche legate al bias algoritmico, è essenziale considerare le distorsioni sistematiche nel modo in cui gli individui interpretano e utilizzano le informazioni. L'ecosistema digitale tende infatti a rafforzare tali distorsioni e i fenomeni di polarizzazione, influenzando sia la disponibilità che l'interpretazione dei dati. In questo contesto, l'automazione rischia di amplificare tali distorsioni, rendendo ancora più urgente la necessità di un'infrastruttura epistemologica solida e di controlli metodologici rigorosi. Per preservare la fiducia nella statistica ufficiale nell'era dell'intelligenza artificiale, trasparenza, alfabetizzazione digitale e integrità della metodologia diventano strumenti imprescindibili (Quattrociocchi, 2023).

La responsabilità nell'uso dell'IA consiste nell'integrare tecnologie IA nei processi statistici in modo etico, trasparente e conforme alle normative, garantendo la qualità dei dati, la protezione della privacy e la fiducia del pubblico. Ciò richiede che tutti i processi siano trasparenti, controllabili e contestabili. I sistemi considerati ad alto rischio nell'ambito statistico dovranno essere progettati garantendo tracciabilità. una supervisione umana attiva e la sicurezza informatica. L'impiego di grandi volumi di dati sensibili impone, inoltre, l'utilizzo di tecniche avanzate di anonimizzazione, misure di sicurezza robuste e piena conformità alla normativa vigente, incluso il GDPR.

L'adozione di una IA responsabile nella statistica ufficiale offre vantaggi significativi, tra cui l'automazione di processi complessi (classificazione di testi, la codifica di risposte aperte nei questionari), l'analisi di grandi volumi di dati (il riconoscimento di im-

magini satellitari) e la costruzione di indicatori innovativi, utili per supportare decisioni pubbliche più tempestive e informate. Tuttavia, ci sono anche aspetti critici, riguardo la qualità e la rappresentatività dei dati detenuti dai privati, la leggibilità degli algoritmi (es., le reti neurali profonde, sono considerati "scatole nere" difficili da interpretare), il rispetto della privacy e la non discriminazione. La trasformazione, dunque, non è solo tecnologica, ma anche culturale e organizzativa e comporta una certa disponibilità di competenze interdisciplinari. Investire in competenze, governance e infrastrutture adeguate consente di coniugare innovazione e rigore, assicurando al contempo qualità, tempestività e affidabilità dell'informazione statistica. In questo senso, l'IA diventa un catalizzatore strategico, capace di rafforzare l'impatto e l'autorevolezza della statistica pubblica.

2.1. Esperienze internazionali

Numerose esperienze internazionali dimostrano come l'impiego dell'intelligenza artificiale stia trasformando profondamente le pratiche statistiche attraverso l'automatizzazione delle procedure analitiche. Si tratta di esperienze che hanno contribuito a consolidare il quadro teorico concettuale dell'approccio responsabile alle nuove tecniche.

L'UNECE High-Level Group for the Modernization of Official Statistics - HLG-MOS ha avviato nel 2019 un progetto pionieristico sul machine learning per evidenziarne il potenziale per migliorare attività complesse come l'editing e l'imputazione dei dati (HLG-MOS, 2023). Sono infatti tali processi a essersi rivelati particolarmente adatti all'automazione e hanno rappresentato una prima concreta opportunità per integrare tecnologie intelligenti nella produzione statistica. L'UNECE ha approfondito il tema nel 2021 pubblicando un rapporto che presenta un framework di qualità volto a supportare il passaggio dalla fase sperimentale all'adozione operativa dei casi d'uso del machine learning realizzati dalle organizzazioni statistiche. Le cinque dimensioni fondamentali individuate - accuratezza, spiegabilità, riproducibilità, tempestività ed economicità - costituiscono una base di riferimento condivisa. Gli ambiti in cui il ML sembra mostrarsi più efficace sono ancora il processo di automazione di editing e imputazione, in cui garantisce coerenza e rapidità, ma anche la codifica e classificazione e l'analisi delle immagini, in cui studi pilota hanno evidenziato livelli di accuratezza superiori al 90% (Stramer, 2020).

Attualmente, l'UNECE sta sviluppando una nuova iniziativa su IA generativa e statistiche ufficiali, con l'obiettivo di esplorare come i Large Language Models (LLM) possano rafforzare l'efficienza dei processi statistici, automatizzare le attività ripetitive e ampliare le capacità di analisi. Il progetto promuove lo sviluppo di infrastrutture adeguate, investimenti nelle competenze, una gestione responsabile dei dati e la creazione di una cultura favorevole all'adozione dell'IA. Viene, inoltre, posta grande attenzione

alla governance e alla gestione del rischio, per garantire che l'innovazione sia condotta in modo trasparente e collaborativo. A tal fine, è stato realizzato un github dove vengono raccolte le più rilevanti esperienze internazionali condotte[3].

Altra importante esperienza europea, avviata nel 2024, è il progetto ESSnet *One stop-shop AI/ML for official statistics* - AIML4OS, sotto il coordinamento dell'Ufficio di statistica irlandese e con la partecipazione di 16 paesi. L'obiettivo è la creazione di una piattaforma condivisa per sviluppare competenze, strumenti e soluzioni IA/ML all'interno della statistica ufficiale. Tra i casi d'uso: l'impiego di dati da osservazione della Terra e immagini satellitari, la classificazione automatica dei testi, l'utilizzo dei LLM per analisi linguistiche, nonché l'automazione di editing e imputazione. Il progetto include anche l'allestimento di un laboratorio tecnico, basato sulla piattaforma open source Onyxia, che consente agli Istituti di statistica europei di accedere a servizi cloud per la lavorazione sicura dei dati. Il laboratorio è supportato da una rete di formazione e da attività di coordinamento infrastrutturale. Nell'ambito di questo progetto, Statistics Netherlands guida un'iniziativa per sviluppare modelli di ML per l'analisi delle reti di supply chain, addestrandoli su un dataset di rete a livello aziendale per il Portogallo, derivato da fonti amministrative. Inoltre, verrà sviluppato un metodo per assegnare pesi ai collegamenti e per integrare le reti con le tabelle input-output dei conti nazionali. La qualità dei modelli sarà verificata applicandoli a diversi paesi e confrontando i risultati con dataset di reti reali. L'obiettivo finale è creare modelli che permettano a tutti gli Istituti nazionali di statistica dell'UE di generare dataset di supply chain aziendali con una qualità di base garantita anche in caso di dati parziali[4].

Anche le capacità previsive incampo economico e sociale traggono beneficio dalle nuove tecniche: l'integrazione tra dati storici e fonti digitali in tempo reale consente di costruire modelli predittivi più sensibili e adattivi, adatti a cogliere evoluzioni rapide nei mercati del lavoro o nei comportamenti sociali. Il nowcasting del PIL è al centro delle sperimentazioni di uso del machine learning del Fondo Monetario Internazionale (Marini, 2023).

Nel campo della produzione di statistiche economiche, a fine 2024, diverse esperienze sono state presentate durante lo Sprint on Artificial Intelligence and Data Science for Economic Statistics del Comitato di esperti su Big Data and Data Science for Official Statistics delle Nazioni unite (UNCEBD, 2024). Si tratta di progetti in produzione che utilizzano le cosiddette *Reproducible analytical pipelines*. Questa strategia, basata su all'uso di strumenti informatici open-source, prevede il riuso delle tecniche di analisi e permette di creare processi statistici verificabili, riducendo la variabilità tra studi.

[3] https://unece.org/statistics/events/GenAI2025 e https://unece.github.io/genAI/

[4] https://cros.ec.europa.eu/book-page/aiml4os-wp11-creating-firm-level-supply-chain-networks-data-using-aiml

Secondo il *Code of practice for statistics* britannico (UK Office for Statistics Regulation, 2022), ciò permette di rafforzare la comparabilità dei dati nel tempo e tra paesi. Tra le attività più interessanti vi è il progetto dell'Istituto nazionale di statistica del Messico (INEGI) che impiega modelli linguistici open-source per automatizzare la codifica delle risposte aperte su occupazione e attività economica nelle principali indagini sulle famiglie. L'iniziativa ha ridotto drasticamente il ricorso alla codifica manuale, migliorando al contempo la coerenza dei risultati. (Pimentel, 2024).

2.2. Focus sulle esperienze dell'Istat

Negli ultimi anni, anche l'Istat ha avviato una serie di studi pilota volti a esplorare l'adozione del machine learning nei processi di editing, imputazione e analisi automatizzata. L'obiettivo è duplice: da un lato, migliorare l'efficienza e la qualità della produzione statistica, dall'altro, garantire che l'innovazione avvenga nel rispetto dei principi metodologici, etici e organizzativi che caratterizzano la statistica ufficiale. Queste attività non rappresentano, quindi, soltanto il segno di avanzamenti sul piano tecnologico, ma si inseriscono in un più ampio percorso di trasformazione organizzativa, rafforzamento delle competenze interne e promozione della collaborazione interdisciplinare.

Le esperienze che seguono illustrano in che modo Istat abbia tradotto nella pratica i principi promossi a livello internazionale sperimentando soluzioni innovative.

Nel primo studio pilota, condotto seguendo le linee guida metodologiche ed etiche dell'UNECE, l'Istat ha applicato strumenti di machine learning per migliorare i processi di editing nell'Anagrafe della Pubblica Amministrazione italiana, utilizzando i dati della banca dati delle Pubbliche Amministrazioni (BDAP) e del Sistema informativo sul funzionamento degli Enti pubblici (SIOPE). I metodi includevano l'identificazione di incongruenze nei dati e l'applicazione di Alberi Decisionali e Foreste Casuali per la classificazione. I risultati hanno dimostrato il potenziale del machine learning per semplificare i processi di editing e imputazione, migliorandone l'accuratezza e l'efficienza (UNECE, 2021). Un secondo studio pilota sull'imputazione dei dati ha riguardato la variabile relativa al livello di istruzione conseguito presente nel Registro di base degli individui. I dati includevano informazioni amministrative del Ministero dell'Istruzione, dell'Università e della Ricerca, dati del censimento del 2011 e dati di indagini campionarie e si voleva valutare i vantaggi dell'introduzione del machine learning rispetto ai modelli statistici classici per risolvere i problemi di imputazione. Per analizzare i dati incompleti di una Regione, sono stati scelti manualmente alcuni parametri e confrontati i risultati, usando diversi algoritmi di apprendimento automatico (come MLP, Random Forest e Log-Linear Model). L'analisi è stata condotta su una piattaforma cloud con Python e Azure, l'accuratezza è stata misurata sia a livello micro che macro. I risultati mostrano che i modelli MLP e log-lineari hanno fornito stime simili. Tuttavia, l'impu-

tazione log-lineare ha funzionato meglio per valori estremi (i.e., i dottorati di ricerca), mentre il modello MLP ha ottenuto una precisione leggermente superiore nel dettaglio micro e non ha richiesto un pretrattamento delle variabili.

Dall'inizio del 2025, l'Istat ricopre un ruolo di rilievo nell'ambito del progetto europeo ESSnet 'One stop-shop AI/ML for official statistics'. L'Istituto è responsabile sia del Work Package 13 sulle tecniche di generazione di dati sintetici, che ha l'obiettivo di bilanciare utilità e riservatezza, sia delle attività di comunicazione e coinvolgimento della comunità di utilizzatori delle nuove tecniche.

Contemporaneamente, sono in corso i primi casi di adozione di intelligenza artificiale in ambito di diffusuione dei dati, sia per la ricerca di informazioni e dati sul sito istituzionale, sia per supportare la risposta alle richieste degli utenti.

Per garantire un'integrazione efficace di queste tecnologie nei processi statistici e organizzativi, è stato fondamentale promuovere una cultura collaborativa che coinvolga statistici, data scientist, professionisti IT ed altri esperti. L'adozione di queste tecniche richiede un coordinamento strutturato tra diverse aree dell'Istituto di statistica, in cui le risorse umane giocano un ruolo centrale nello sviluppo delle competenze del personale e nell'acquisizione di nuovi talenti, assicurando che le capacità interne siano sempre aggiornate. Parallelamente, il settore IT deve garantire infrastrutture adeguate a supportare l'implementazione di strumenti avanzati, favorendo un ecosistema tecnologico affidabile e scalabile. Occorre inoltre tener conto delle implicazioni normative e regolamentarie, motivo per cui il settore legale gioca un ruolo essenziale nel garantire il rispetto delle norme etiche e di tutela dei dati. Infine, le relazioni internazionali favoriscono il dialogo tra istituzioni, permettendo lo scambio di esperienze e la definizione di soluzioni condivise sui temi emergenti (De Cubellis, Grippo, 2024).

Per consolidare questa cultura collaborativa, l'Istat investe nella formazione del personale, organizzando corsi specifici e promuovendo la partecipazione a programmi europei di alta specializzazione, come il Master europeo in Statistica ufficiale e l'European statistical training programme, oltre ai corsi offerti dalla Scuola nazionale di Pubblica Amministrazione e dalle Università. Inoltre, l'Istituto ha avviato programmi di tirocinio che consenteno agli studenti universitari di sviluppare tesi e progetti basati su tecniche di machine learning direttamente in un contesto applicativo, rafforzando così il legame tra ricerca accademica e innovazione operativa.

3. Una governance per un'implementazione responsabile delle tecniche di IA/ML

Si è visto come l'adozione di queste tecniche sollevi questioni cruciali legate all'etica e alla sicurezza. La sfida per gli Istituti nazionali di statistica non è solo applicare tecnologie intelligenti, ma predisporre strutture di governance che siano in grado di controllarne, valutarne e indirizzarne gli impieghi. Questo significa prevedere ex ante

i rischi di opacità, distorsione o discriminazione, rafforzando capacità istituzionali di auditing e partecipazione, perché l'attuale quadro tradizionale di assicurazione della qualità della statistica ufficiale non appare più sufficiente per includere anche processi di IA.

Per affrontare la complessità degli algoritmi spesso opachi, è fondamentale promuovere l'adozione di modelli interpretabili, documentare metodologie e dati in modo accurato e aderire a standard internazionali come quelli previsti dal quadro normativo europeo. Quest'ultimo, articolato nella Legge sull'IA, nel Regolamento (UE) 2024/1689 e nel Patto sull'IA, stabilisce regole armonizzate per garantire un uso affidabile e responsabile dell'IA, prevedendo classificazioni per livelli di rischio, sanzioni significative per la non conformità e strumenti innovativi come sandbox regolamentari.

Parallelamente, è essenziale mitigare il rischio di bias e discriminazioni nei sistemi intelligenti, poiché gli algoritmi apprendono da dati che possono riflettere pregiudizi preesistenti. Per garantire equità, gli istituti statistici devono condurre controlli rigorosi sulla qualità e sulla rappresentatività dei dati, impiegare tecniche di debiasing e monitorare costantemente i risultati. La gestione di grandi volumi di informazioni sensibili rende, inoltre, imprescindibile l'adozione di strumenti avanzati per la protezione dei dati, come l'anonimizzazione, la pseudonimizzazione e protocolli di sicurezza conformi al GDPR per bilanciare l'accesso ai dati con la tutela dei diritti dei cittadini

Nonostante l'efficienza introdotta dall'automazione, il ruolo della supervisione umana resta centrale. È quindi necessario istituire comitati etici, definire linee guida chiare e garantire che l'intervento umano sia sempre possibile nelle decisioni critiche. In questo contesto, la governance non deve essere percepita come un vincolo, bensì come un motore strategico capace di facilitare la transizione verso una gestione più moderna della statistica ufficiale, grazie anche all'uso mirato di tecnologie leggere, alla gestione attenta del cambiamento e allo sviluppo delle competenze.

La trasformazione richiede un ripensamento organizzativo delle strutture interne, con l'adozione di approcci innovativi come il *Data Mesh*, che decentralizza la gestione dei dati, e strategie di *open science*, che promuovono la collaborazione tra attori pubblici e privati. La rapida evoluzione tecnologica rende prioritario investire in formazione continua, favorire lo scambio di competenze tra esperti e attrarre nuovi talenti nel campo della data science, affinché il capitale umano sia pronto ad affrontare le sfide emergenti.

A supporto di questa trasformazione, la strategia delle *Reproducible analytical pipelines* (RAP) per la produzione statistica contribuisce alla mitigazione dei bias attraverso una gestione rigorosa dei dati, promuove la tracciabilità tramite strumenti di versionamento come Git, e migliora la qualità dei risultati automatizzando e standardizzando i processi analitici. Le RAP possono diventare, quindi, lo strumento per realizzare un'IA

responsabile che consenta agli Istituti statistici di implementare soluzioni affidabili, trasparenti ed eque, rafforzando la fiducia del pubblico e allineando l'innovazione tecnologica ai valori dell'integrità e dell'etica.

4. Conclusioni

L'utilizzo integrato di dati ufficiali e di fonti non tradizionali – come immagini satellitari, segnali da sensori, testi non strutturati o dati da reti sociali – richiede lo sviluppo di quadri metodologici e normativi in grado di garantire qualità, rappresentatività e trasparenza. Questa integrazione è una delle principali opportunità offerte dall'IA; essa può ridurre i costi della statistica pubblica e gli oneri sui rispondenti, migliorando al contempo la granularità e la tempestività delle statistiche.

L'adozione di queste tecnologie da parte delle istituzioni statistiche rappresenta, quindi, una leva cruciale per migliorare l'efficienza produttiva e la capacità di risposta alle esigenze di una società sempre più dinamica e interconnessa. Tuttavia, questa trasformazione non è soltanto tecnologica, essa implica una sfida epistemologica, etica e organizzativa che interessa l'intero ciclo di vita della statistica, dalla raccolta alla diffusione e all'uso dei dati da parte della collettività (Chelli, 2024).

Le esperienze di sperimentazioni nazionali e internazionali riportate non sono semplici esempi operativi, ma manifestazioni coerenti del quadro concettuale di adozione di una IA responsabile guidata da una visione strategica che integra innovazione tecnologica, responsabilità istituzionale, per costruire un sistema statistico più efficace e inclusivo.

In questo scenario, le istituzioni statistiche europee sono chiamate a svolgere un ruolo guida nella costruzione di un'infrastruttura statistica in cui l'IA non sostituisca il metodo, ma ne potenzi la capacità analitica e predittiva. Ciò implica la promozione di un modello di governance cooperativa, basato su standard comuni, interoperabilità e condivisione delle conoscenze.

Per rendere operativa l'integrazione dell'IA nei processi statistici, occorre realizzare alcune azioni concrete. Tra queste l'utilizzo sempre più esteso di laboratori sperimentali inter-istituzionali (si vedano gli accordi inter-istituzionali per l'utilizzo del super calcolatore del Cineca) per testare algoritmi su dati reali, con protocolli condivisi di validazione, documentazione e auditing. Così come, l'adozione di strategie di *open science*, che promuovono la collaborazione tra attori pubblici e privati. Poi, il favorire il riuso, il benchmarking e l'interoperabilità dei dati su piattaforme comuni (è il caso dei dataset messi a disposizione degli istituti dei paesi membri da Eurostat sulla propria Piattaforma Big data). Certamente, l'adozione di strumenti di explainable AI (XAI) e lo sviluppo di *Reproducible analytical pipelines* (RAP) per la produzione statistica per migliorare gli output automatizzando e standardizzando i processi. Infine, l'implementazione di siste-

mi di monitoraggio continuo a livello europeo e nazionale per valutare l'impatto dell'IA su qualità, equità e fiducia pubblica.

L'Istat dovrà, inoltre, rafforzare la propria funzione di guida nel Sistema Statistico Nazionale, promuovendo reti collaborative, sostenendo la costruzione di una legge quadro aggiornata sull'uso dell'IA nella statistica ufficiale e rendendo il Sistema un punto di riferimento per la misurazione dei fenomeni emergenti della società digitale.

In definitiva, l'intelligenza artificiale non è un fine in sé, ma uno strumento per valorizzare il sapere umano e statistico, ponendolo al servizio del bene collettivo e del progresso delle politiche pubbliche basate su dati di qualità. Solo attraverso un'integrazione responsabile, trasparente e partecipata sarà possibile costruire una statistica ufficiale più flessibile, personalizzata e capace di rispondere alle sfide del nostro tempo.

Bibliografia

Chelli, F.M. (2024). Intervento inaugurale, *XV Conferenza Nazionale di Statistica*, Roma, scaricato da https://www.istat.it/wp-content/uploads/2024/06/Saluti-istituzionali-del-Presidente-Francesco-Maria-Chelli.pdf

De Cubellis, M. & Grippo, G. (2024). Organizational Sustainability to Support Trusted Smart Statistics: Istat's Experience. *Statistical Journal of The Iaos*. Jan, 1-14, *ISSN 1874-7655*

High-Level Group for the Modernisation of Official Statistics (HLG-MOS) (2023), Large Language Models for Official Statistics. HLG-MOS White Paper. Modernstats HLG-MOS

Loru, E., Nudo, J., Di Marco, N., Santirocchi A., Atzeni, R., Cinelli, M., Cestari, V., Clelia Rossi-Arnaud, C. and W. Quattrociocchi (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences (PNAS),* 122 (42) e2518443122 https://doi.org/10.1073/pnas.2518443122

Marini, M. (2023). Responsible AI for data and statistics at the IMF, *presented at the 64th ISI World Statistics Congress*, Ottawa, Canada.

Pimentel A. (2024), Leveraging Open-Source Language Models for Automatic Codification of Employment and Economic Activity in Household Surveys, presented *at UNCEBD SPRINT ON Artificial Intelligence and Data Science for Economic Statistics*. Retrived from: https://unstats.un.org/bigdata/events/2025/ai-data-science/webinar2/presentations/Alejandro%20Pimentel%20-%20sprintAI2024.pdf

Quattrociocchi, W., Scala A., & Sunstein, C. R. (2016). Echo chambers on Facebook, Retrived from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110.

Stramer, K. (2020). UNECE – HLG-MOS Machine Learning Project Classification and Coding Theme *Report*, Retrived from: https://unece.org/sites/default/files/2022-01/ML_06_Classification%20and%20Coding%20Theme%20Report.pdf

UK Office for Statistics Regulation (2022). Using Reproducible Analytical Pipelines (RAP) to improve statistics, *Code of practices for statistics*, Retrived from: https://code.statisticsauthority.gov.uk/case-studies/using-reproducible-analytical-pipelines-rap-to-improve-statistics/

UNCEBD (2024). *Proceedings of the Sprint on Artificial Intelligence and Data Science for Economic Statistics*, Webinar nov. - dic., Retrived from: https://unstats.un.org/bigdata/events/2025/ai-data-science/

UNECE (2021). *Machine Learning for Official Statistics*, Geneva, UN

UNECE (2024). Follow-up on the Conference of European Statisticians seminar "Data ethics – a key enabler *of social acceptability*, Geneva, 20 - 21 June, Item 8, ECE/CES/2024/19. Retrived from: https://unece.org/statistics/documents/2024/06/working-documents/follow-conference-european-statisticians-seminar

PRINTED IN FEBRUARY 2026
ON BEHALF OF
GIAPETO EDITORE

www.giapeto.it