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

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.

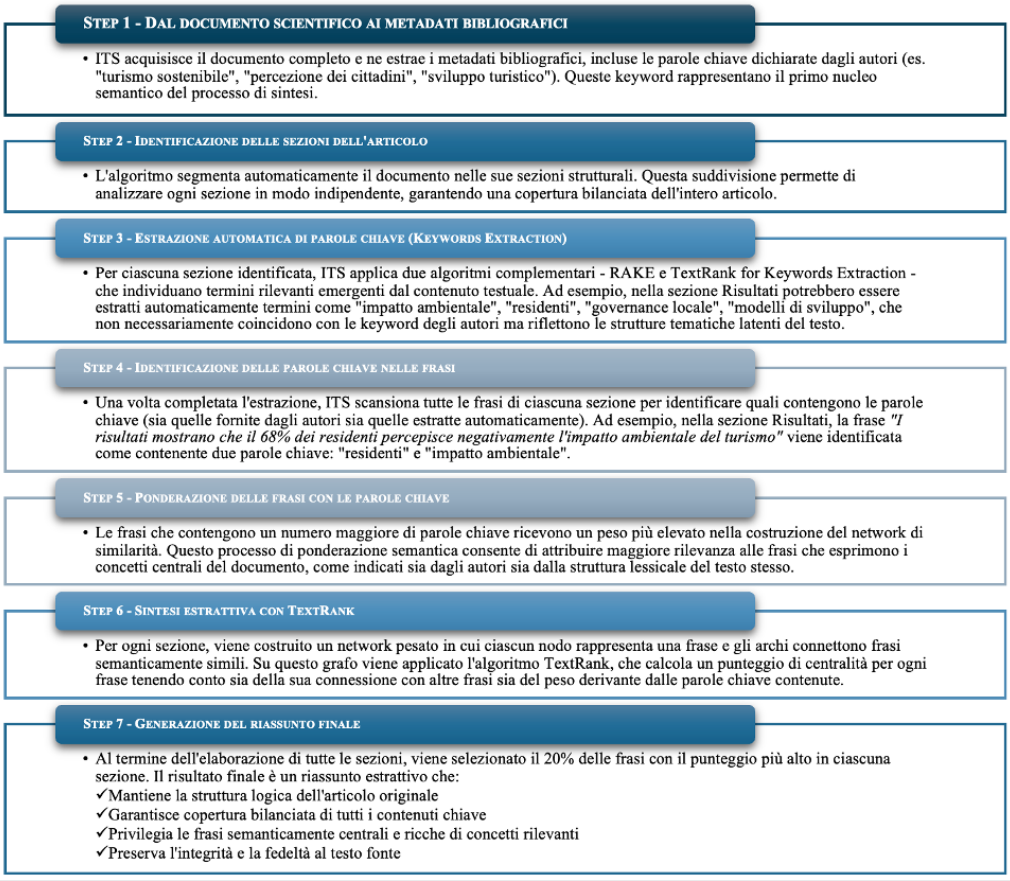
Combinando i due algoritmi di estrazione, l'ITS genera un insieme ampliato e articolato di parole chiave, in grado di rappresentare sia le dimensioni semantiche esplicite

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

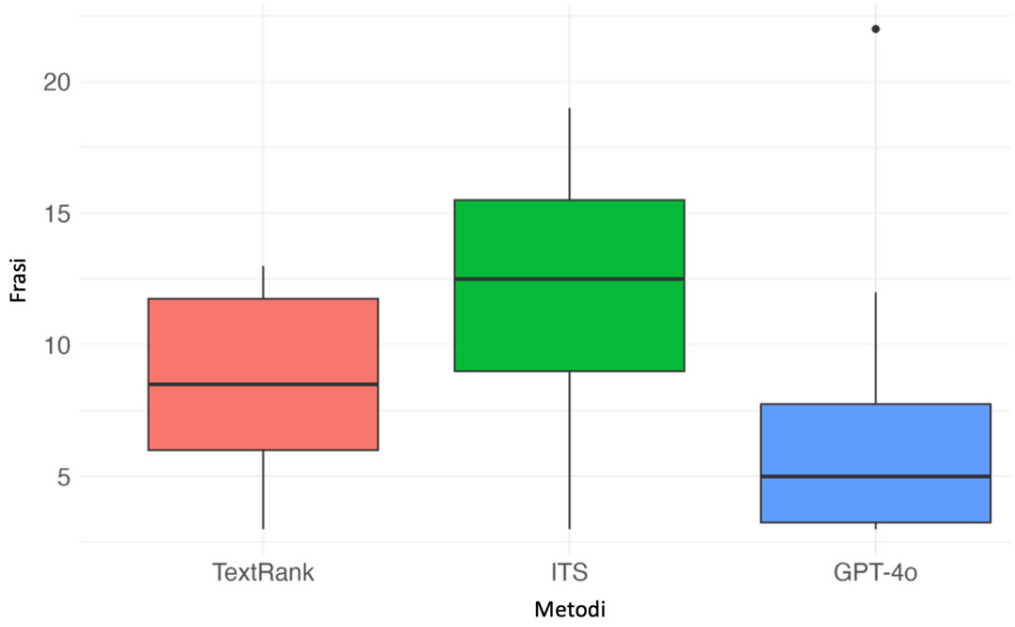
I risultati dell'analisi statistica (Tabella 3) forniscono un quadro chiaro delle relazioni tra i metodi:

- **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.

PRINTED IN FEBRUARY 2026
ON BEHALF OF
GIAPETO EDITORE

www.giapeto.it