



This document is designed for use in Google Docs, Microsoft Office, Libre Office, or other similar text editing software. Find PDF and Markdown versions of this

document in our Github repository or on our website.

MDPI Dataset	Il dataset è composto da 10 articoli in formato pdf estratti tramite web scraping su python dal sito web MDPI, un editore di riviste scientifiche. In particolare gli articoli fanno parte del Volume 12, Issue 1 della sezione “Mathematics”, ovvero sono i primi 10 articoli pubblicati nel mese di gennaio 2024 che trattano argomenti matematici. Questo dataset verrà utilizzato per analizzare gli argomenti trattati negli articoli tramite un approccio topic modelling.
DATASET LINK https://www.mdpi.com/2227-7390/12/1	DATA CARD AUTHOR(S) Valerio Valentini: (Owner, Contributor, Manager)

Authorship		
Publishers		
PUBLISHING ORGANIZATION(S)	INDUSTRY TYPE(S)	CONTACT DETAIL(S)
MDPI	Academic - Scientific	Contact: +41 61 683 77 34 Website: https://www.mdpi.com/2227-7390/12/1
Dataset Owners		
TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)

Valerio Valentini	<p>Dataset Owner(s): Valerio Valentini</p> <p>Affiliation: Università LUMSA</p> <p>Contact: v.valentini4@lumsastud.it</p>	<p>Haoxiang Zhang, Chao Liu, Jianguang Ma and Hui Sun</p> <p>Mikhail Malykh, Mark Gambaryan , Oleg Kroytor and Alexander Zorin</p> <p>Jia Zhao and Yu Qiao’</p> <p>Anatoly Nikolaevich Filippov</p> <p>Xulin Ma, Jiajia Tan, Linan Zhu , Xiaoran Yan and Xiangjie Kong</p> <p>Junchao Zhang, Jun Huang and Changjie Li</p> <p>Haiyan Su and Linlin Chen</p> <p>Jing Zhou, Lang Liu, Yuan Zhao, Mengbo Zhu, Ruofan Wang and Dengdeng Zhuang</p> <p>Yuri V. Tyutyunov</p> <p>Vasily N. Govorukhin and Vyacheslav G. Tsybulin</p> <p>Shi-Wei Wang, Guang-Xin Huang and Feng Yin</p>
-------------------	--	--

Dataset Overview		
DATA SUBJECT(S)	DATASET SNAPSHOT	CONTENT DESCRIPTION
Data about mathematical topics and theories	<p>Size of Dataset:</p> <p>4.7 MB</p> <p>371 KB</p> <p>351 B</p> <p>2.2 MB</p> <p>698 KB</p> <p>381 KB</p> <p>257 KB</p> <p>11.2 MB</p> <p>13 MB</p> <p>958 KB</p> <p>Above: Dimensione degli articoli, dal dal primo al decimo e ultimo</p>	<p>Additional Notes: Il dataset è composto da 10 articoli in formato testuale per i quali non è possibile individuare istanze e classi e e non è quindi nemmeno possibile eseguire analisi statistiche su queste ultime</p>
Sensitivity of Data		
SENSITIVITY TYPE(S)	FIELD(S) WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING
Mathematical Data	Non è stato raccolto alcun tipo di dato sensibile, intenzionalmente o meno	-
RISK TYPE(S)	SUPPLEMENTAL LINK(S)	RISK(S) AND MITIGATION(S)
No Known Risks	-	-
Dataset Version and Maintenance		
MAINTENANCE STATUS	VERSION DETAILS	MAINTENANCE PLAN

Limited Maintenance

(The data will not be updated, but any technical issues will be addressed.)

Current Version: 1.0

Last Updated: 05/2024

Release Date: 01/2024

Il dataset verrà modificato o corretto solo durante la fase di analisi e solo se viene rilevata la presenza di documenti corrotti o di errori nei testi

Versioning: il dataset verrà versionato in caso di modifiche durante la fase di analisi

Updates: il dataset verrà aggiornato durante la fase di analisi solamente per correggere eventuali errori presenti

Errors: un errore può consistere nella corruzione di un articolo nella fase di download durante il web scraping o nel danneggiamento del contenuto di un articolo durante la fase di analisi

Feedback: Il feedback sul dataset verrà fornito direttamente da chi utilizzerà il dataset per le sue analisi

Example of Data Points

PRIMARY DATA MODALITY	SAMPLING OF DATA POINTS	DATA FIELDS
Text Data	No sampling (full data)	Non ci sono data fields, per ogni articolo viene preso in considerazione tutto il testo presente
TYPICAL DATA POINT		A TYPICAL DATA POINT
Il testo a destra rappresenta una porzione di testo estratto dal primo articolo del dataset		Ship targets are crucial combat units in modern maritime warfare, and it is important to recognize them accurately and credibly to enhance maritime situational awareness and gain an advantage. Infrared (IR) imaging technology is advantageous due to its all-weather capability, long-range perception, and strong concealment. It, along with visible light and synthetic aperture radar (SAR) imaging, forms an important means of acquiring feature information about ship targets and is widely used in ship automatic target recognition (SATR) tasks [1–3]. In the marine environment, different types of ships exhibit varying degrees of thermal characteristics and IR radiation spectra within the IR spectrum. This paper focuses on ship IR automatic target recognition (SIATR) technology, which utilizes sensors to capture IR images of ship targets. By combining image processing, recognition algorithms, and other techniques, this technology automatically and accurately extracts shape, IR radiation, and other feature information from the targets.

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Web Scraping	<p>Source: https://www.mdpi.com/2227-7390/12/1</p> <p>Platform: Jupyter Notebook, Python 3.9</p> <p>Is this source considered sensitive or high-risk? No</p> <p>Dates of Collection: [MARZO 2024 - MAGGIO 2024]</p> <p>Primary modality of collected data: Text Data</p> <p>Update Frequency for collected data: Solo se vengono rilevati errori</p>	<p>https://www.mdpi.com/2227-7390/12/1/168/pdf?version=1704382967</p> <p>https://www.mdpi.com/2227-7390/12/1/167/pdf?version=1704418174</p> <p>https://www.mdpi.com/2227-7390/12/1/166/pdf?version=1704700168</p> <p>https://www.mdpi.com/2227-7390/12/1/165/pdf?version=1704442425</p> <p>https://www.mdpi.com/2227-7390/12/1/164/pdf?version=1704369944</p> <p>https://www.mdpi.com/2227-7390/12/1/163/pdf?version=1704688433</p> <p>https://www.mdpi.com/2227-7390/12/1/162/pdf?version=1704357907</p> <p>https://www.mdpi.com/2227-7390/12/1/161/pdf?version=1704351241</p> <p>https://www.mdpi.com/2227-7390/12/1/160/pdf?version=1704351673</p> <p>https://www.mdpi.com/2227-7390/12/1/159/pdf?version=1704294566</p> <p>Additional Notes: Sono riportati i link tramite i quali sono stati scaricati i 10 articoli dal sito web</p>

Motivations & Intentions		
Motivations		
PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
Mining	Topic modelling	<i>Estrarre gli argomenti di cui trattano gli articoli e analizzare il contributo di ogni parola al topic a cui appartiene</i>
Intended Use		
DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
Safe for research use	[Suitable Use Case] : estrazione della miscela di topic da ogni articolo per eseguire topic modelling, così come document clustering o classification	<p>[Unsuitable Use Case] : Il dataset non è adatto a eseguire analisi del sentiment in quanto contiene articoli matematici con tematiche affrontate oggettivamente. Non è quindi opportuno o utile eseguirvi Sentiment Analysis</p> <p>[Unsuitable Use Case]: il dataset non è adatto a eseguire Opinion Mining in quanto, come riportato sopra, non sono contenute opinioni soggettive negli articoli, ma solo concetti matematici oggettivi</p>
	RESEARCH AND PROBLEM SPACE(S)	CITATION GUIDELINES
	Il dataset ha l'obiettivo di poter essere analizzato per eseguire topic modelling, quindi per estrarre gli argomenti di cui tratta ciascun articolo	<p>Guidelines & Steps: Al momento della citazione di questo dataset dovrà essere riportato anche il suo creatore, i suoi intended uses, i mezzi impiegati per il suo ottenimento e altre informazioni presenti nel testo in formato json riportato di seguito</p> <p>JSON:</p> <pre>{ "autore": "Valerio Valentini", "titolo": "MDPI GEN 2024 Dataset", "numero di articoli": 10, "anno": "2024", "intended uses": "topic modeling", "document clustering", "document classification" "libreria web scraping": "Selenium", "BeautifulSoup" }</pre>

Access, Retention, & Wipeout		
Access		
ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
External - Open Access	GitHub	<i>Questo dataset non richiede prerequisiti per accedervi</i>
	POLICY LINK(S)	ACCESS CONTROL LIST(S)
	GitHub	Questo dataset non prevede liste di controllo per gli accessi
Retention		
	DURATION	POLICY SUMMARY
	Il dataset può essere conservato senza limiti temporali	Retention Plan ID: RP-2024-01 Summary: Il dataset sarà conservato in modo permanente, senza una data di scadenza specifica per la sua eliminazione in modo da mantenere i dati accessibili e disponibili a tempo indeterminato per scopi di ricerca, analisi e riferimento futuri
	PROCESS GUIDE	EXCEPTION(S) AND EXEMPTION(S)
	<i>Conservare il dataset in un ambiente stabile e compatibile, ricordarsi di fare sempre una copia prima di eseguire le analisi e di mantenere la versione originale intatta, controllare periodicamente la presenza di nuove versioni del dataset in caso di identificazione di errori</i>	Summary: nel caso in cui il dataset scaricato venisse alterato senza possibilità di ripristinare le modifiche è possibile ricaricarlo dalla fonte utilizzata inizialmente
Wipeout and Deletion		
	DURATION	DELETION EVENT SUMMARY
	Non c'è una data precisa nella quale deve essere eseguita l'eliminazione del dataset scaricato	Sequence of deletion and processing events: Identificare i dati da eliminare e le conseguenze della loro eliminazione sull'analisi svolta Utilizzo di procedure di eliminazione sicure e irreversibili per garantire la distruzione completa dei dati

	ACCEPTABLE MEANS OF DELETION	POST-DELETION OBLIGATIONS
	<p>Cancellazione definitiva dei file dai sistemi di archiviazione impiegati</p> <p>Distruzione fisica dei supporti di archiviazione</p>	<p>Sequence of post-deletion obligations:</p> <p>Verificare la corretta esecuzione dell'eliminazione dei dati e della distruzione completa e irreversibile dei dati eliminati</p>
	OPERATIONAL REQUIREMENT(S)	EXCEPTIONS AND EXEMPTIONS
	<p>Wipeout Integration Operational Requirements:</p> <p>Assicurarsi che il software di wipeout sia integrato correttamente nei sistemi operativi e nelle piattaforme utilizzate dall'organizzazione</p>	<p>Policy Exception bug: eliminazione non corretta</p> <p>Summary: nel caso in cui avvenga un'eliminazione non corretta del dataset si deve procedere analizzando il bug che ha portato a quella problematica per poi ritentare l'eliminazione definitiva del dataset</p>

Provenance		
Collection		
METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Scraped or Crawled	<p>Source: https://www.mdpi.com/2227-7390/12/1</p> <p>Platform: Jupyter Notebook, Python 3,9</p> <p>Is this source considered sensitive or high-risk? No</p> <p>Dates of Collection: [03/2024 - 05/2024]</p> <p>Primary modality of collected data: Text Data</p> <p>Update Frequency for collected data: Static</p>	<p>https://www.mdpi.com/2227-7390/12/1/168/pdf?version=1704382967 https://www.mdpi.com/2227-7390/12/1/167/pdf?version=1704418174 https://www.mdpi.com/2227-7390/12/1/166/pdf?version=1704700168 https://www.mdpi.com/2227-7390/12/1/165/pdf?version=1704442425 https://www.mdpi.com/2227-7390/12/1/164/pdf?version=1704369944 https://www.mdpi.com/2227-7390/12/1/163/pdf?version=1704688433 https://www.mdpi.com/2227-7390/12/1/162/pdf?version=1704357907 https://www.mdpi.com/2227-7390/12/1/161/pdf?version=1704351241 https://www.mdpi.com/2227-7390/12/1/160/pdf?version=1704351673 https://www.mdpi.com/2227-7390/12/1/159/pdf?version=1704294566</p> <p>Additional Notes: Sono riportati i link tramite i quali sono stati scaricati i 10 articoli dal sito web</p>
COLLECTION CADENCE	DATA INTEGRATION	DATA PROCESSING
<p>Static</p> <p>(Data was collected once from single or multiple sources.)</p>	Non ci sono data fields, per ogni articolo viene preso in considerazione tutto il testo presente	<p>Description: è stato impiegato web scarping per scaricare i 10 articoli del dataset</p> <p>Methods employed: web scraping</p> <p>Tools or libraries: Selenium, BeautifulSoup</p>
Collection Criteria		
DATA SELECTION	DATA INCLUSION	DATA EXCLUSION
Sono stati selezionati solamente i primi 10 articoli dalla raccolta di gennaio 2024 in quanto ritenuti in numero sufficiente per eseguire topic modeling	Per ogni articolo è stato considerato l'intero testo	Nessuna parte dell'articolo è stata eliminata per la creazione del dataset

Relationship to Source		
USE & UTILITY(IES)	BENEFIT AND VALUE(S)	LIMITATION(S) AND TRADE-OFF(S)
Il dataset contiene 10 articoli matematici che vengono utilizzati per eseguire topic modeling in quanto contengono molteplici argomenti espressi in maniera oggettiva	Il principale benefici che si può trarre dal dataset è la possibilità di eseguirvi diversi tipi di analisi, come topic modeling, document classification, document clustering, sia per fini didattici sia per fini informativi	Tra le limitazioni principali potrebbe essere considerato il numero ridotto di articoli contenuti nel dataset (10), il che potrebbe limitare alcuni tipi di analisi.

Extended Use		
Use with Other Data		
SAFETY LEVEL	KNOWN SAFE DATASET(S) OR DATA TYPE(S)	BEST PRACTICES
Conditionally safe to use with other datasets	<p>Dataset or Data Type: qualsiasi dataset che sia composto esclusivamente da documenti di testo, preferibilmente in formato pdf</p> <p>Dataset or Data Type: un dataset che include altri articoli matematici scaricati sempre dal sito web di MDPI</p>	Per utilizzare questo dataset in unione ad altri dataset esterni è buona norma accertarsi che la lunghezza dei file di testo sia pressoché simile, per evitare sbilanciamenti in eventuali analisi di frequenza sui termini, così come è importante accertarsi che il dataset esterno contenga file di testo scritti nella stessa lingua del dataset di cui stiamo trattando (quindi in inglese)
	KNOWN UNSAFE DATASET(S) OR DATA TYPE(S)	LIMITATION(S) AND RECOMMENDATION(S)
	<p>Dataset or Data Type: è assolutamente sconsigliato unire al dataset originale un dataset che contiene dati in forma tabellare o in generale dati che non sono in formato testuale</p>	Se si utilizza un dataset con un formato di dati diversi o incompatibile con il dataset preso in esame, potrebbe essere necessario un lavoro aggiuntivo per l'armonizzazione dei dati, aumentando la probabilità che emergano errori, e anche che i risultati delle analisi non siano validi
Forking & Sampling		
SAFETY LEVEL	ACCEPTABLE SAMPLING METHOD(S)	BEST PRACTICE(S)
Conditionally safe to fork and/or sample	Cluster Sampling Random Sampling Weighted Sampling	Tra le best practice per campionare il dataset troviamo prima di tutto la scelta del metodo di campionamento, scelta che deve essere eseguita in base alla tipologia di analisi che si intende eseguire. Può essere inoltre fondamentale valutare la rappresentatività del campione, per verificare che non si sia perduto un numero eccessivo di informazioni durante il sampling, per questo è anche importante determinare accuratamente la dimensione del campione che si vuole ottenere.
	RISK(S) AND MITIGATION(S)	LIMITATION(S) AND RECOMMENDATION(S)

	Il rischio fondamentale a cui si potrebbe andare incontro a seguito di sampling sul dataset preso in esame è legato alla sua dimensione contenuta (solo 10 articoli). Dunque, anche i campioni estratti saranno di dimensioni ridotte rischiando, nel caso in cui il campione sia eccessivamente piccolo, di perdere un numero eccessivo di informazioni e di perdere quindi la rappresentatività del campione	La dimensione ridotta del dataset pone una serie di limitazioni, riguardanti principalmente il campione estratto. Quest’ultimo infatti potrebbe avere una limitata potenza statistica e quindi potrebbe non essere in grado di rappresentare completamente la variabilità dei dati.				
Use in ML or AI Systems						
DATASET USE(S)	NOTABLE FEATURE(S)	USAGE GUIDELINE(S)				
Training Testing	Dato che il dataset è composto da file testuali non è possibile individuare distribuzioni di features o relazioni tra istanze	Alcuni degli usi per cui potrebbe essere impiegato questo dataset in ambito di ML sono: Addestrare e testare un modello per classificare gli articoli Addestrare e testare un modello per estrarre entità come nomi di persone, luoghi, organizzazioni, ecc., dai testi Addestrare e testare un modello per tradurre automaticamente il testo da una lingua all’altra. Addestrare e testare un modello per generare nuovi testi basandosi sui testi del dataset				
	DISTRIBUTION(S)	KNOWN CORRELATION(S)				
	<table><tr><td>Train</td><td>80</td></tr><tr><td>Test</td><td>20</td></tr></table> Additional Notes: Dato che il dataset ha dimensioni ridotte si consiglia di scegliere uno splitting tra dataset di training e testing che assegni gran parte dei testi alla fase di addestramento, in modo da riuscire a cogliere al massimo la variabilità dei dati. È inoltre consigliato eseguire cross validation, sempre a causa delle dimensioni del dataset	Train	80	Test	20	Dato che il dataset è composto da file testuali non è possibile individuare correlazioni tra le features
Train	80					
Test	20					

Transformations

Synopsis

TRANSFORMATION(S) APPLIED	FIELD(S) TRANSFORMED	LIBRARY(IES) AND METHOD(S) USED
Web scraping	Numero di articoli scaricati rispetto al numero totale di articoli presenti	<p>Method: è stato impiegato web scraping per estrarre dal sito web di MDPI i primi 10 articoli di gennaio 2024 dalla sezione Mathematics.</p> <p>Platforms, tools, or libraries:</p> <p>Selenium per automatizzare l’interazione con il browser web e scaricare la pagina contenente gli articoli. Nel particolare esegue l'apertura della pagina web del sito MDPI, accetta i cookie, estrae il codice HTML della pagina e quindi chiude il browser</p> <p>BeautifulSoup per analizzare il codice HTML della pagina web precedentemente estratta attraverso Selenium e individuare i tag dei singoli articoli pdf per poi scaricarli tramite i loro link e salvarli sul disco per formare il dataset</p> <p>Transformation Results:</p> <p>Come risultato è stato ottenuto il dataset composto da 10 articoli in formato pdf</p>

Annotations & Labeling

ANNOTATION WORKFORCE TYPE	ANNOTATION CHARACTERISTIC(S)	ANNOTATION DESCRIPTION(S)
Code Annotations	<p>Gli articoli del dataset sono stati rinominati nel seguente modo:</p> <p>Articolo_1, Articolo_2, Articolo_3, Articolo_4, Articolo_5, Articolo_6, Articolo_7, Articolo_8, Articolo_9, Articolo_10,</p>	<p>Durante la fase di web scraping, in particolare dopo aver scaricato gli articoli tramite BeautifulSoup, ognuno di loro è stato aperto e salvato sul disco con un nuovo nome, corrispondente all'ordine di download del file. Questo rende possibile un'immediata individuazione dello specifico articolo durante le analisi successive.</p>

Validation Types

METHOD(S)	BREAKDOWN(S)	DESCRIPTION(S)
Validazione umana della dimensione degli articoli	La valutazione ha riguardato la dimensione dei singoli articoli così come il confronto tra loro dimensioni	<p>Method: Dopo che i testi sono stati estratti tramite web scraping, sono stati sottoposti a revisione umana per verificare che contenessero sufficienti informazioni per svolgervi in seguito un’analisi</p> <p>Validation Results: Dopo la valutazione è stato possibile accertarsi che gli articoli pdf scaricati tramite web scarping sono sufficientemente informativi per il tipo di analisi che vi verrà eseguita</p> <p>Additional Notes: È stato anche verificato che le dimensioni degli articoli fossero simili in modo da non alterare successive analisi statistiche (per esempio analisi di frequenza sulle singole parole)</p>

Sampling Methods

METHOD(S) USED	CHARACTERISTIC(S)	SAMPLING CRITERIA
La libreria BeautifulSoup è stata utilizzata per individuare e estrarre solamente i primi 10 articoli dal codice html scaricato precedentemente tramite Selenium	Il sito web di MDPI nella sezione Mathematics di gennaio 2024 conteneva 168 articoli. Ai fini dell'analisi svolta con questo dataset, sono stati estratti solamente i primi 10 articoli	La scelta di estrarre solamente 10 articoli piuttosto che estrarli tutti (168) è legata a un fattore di convenienza, ovvero per semplificare computazionalmente l'analisi e per mantenere il numero di argomenti estratti tramite topic modeling

Known Applications & Benchmarks

ML APPLICATION(S)	EVALUATION RESULT(S)	EVALUATION PROCESS(ES)
<i>Topic Modeling</i>	(LDA) Model Card: Coerenze dei 7 topic: -1.70, -0.83, -2.13, -0.51, -0.43, -0.55, -0.84 (BTM) Model Card: Coerenza: -2.57, -14.39, -38.81, -34.42, -24.63, -3.41, -67.53	Sia per il modello LDA che per il Modello BTM è stata valutata la coerenza di ogni topic tramite metodo UMass
	DESCRIPTION(S) AND STATISTIC(S)	EXPECTED PERFORMANCE AND KNOWN CAVEATS
	LDA e BTM I modelli LDA e BTM sono stato creati, addestrati (adottando come parametri sia il corpus dei documenti sia il vocabolario ottenuto sempre dai 10 articoli del dataset) e infine valutati tramite metodo Umass	LDA e BTM Sia per il modello LDA che per il modello BTM è stata riportata la miscela di topic (con il relativo peso) per ogni documento, così come la lista (pesata nel caso del modello LDA) delle parole che hanno contribuito maggiormente a identificare ogni topic



The [Data Cards Playbook ↗](#) by Google Research is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

You are free to share and adapt this work under the [appropriate license terms ↗](#).