HOMEWORK 5

ALESSIO MARINUCCI

RICCARDO FELICI

https://github.com/alemari7/Hw5_IDD

OBIETTIVO DELLO STUDIO

- **Analisi delle sorgenti di dati**
- **Definizione di uno schema mediato con almeno 20 attributi**
- **Allineamento dello schema mediato utilizzando una soluzione con ChatGPT**
- **Popolamento dello schema mediato**
- **Calcolo del Record Linkage e Pairwise Matching**
  - **Creazione di una ground-truth con almeno 100 coppie**
  - **Definizione e testing di almeno due strategie di blocking**
  - **Definizione di due diverse strategie di pairwise matching**
  - **Confronto Risultati**

Data Analysis

Mediated Schema

Ground-truth

Blocking

Pairwise matching

Risultati

PIPELINE

DATA ANALYSIS

Estrazione file in diversi formati a partire dalla cartella *homework*

Per ogni sottocartella, decompressione file zip, creando la cartella *extracted_files*

Utilizzo di una funzione **'json_to_csv'** per convertire tutti i file in formato csv, più facilmente gestibili

MEDIATED SCHEMA

- **Utilizzo di Python per definire tutti gli attributi presenti nei dataset, mostrando la lista dei file in cui si usano quegli attributi, come si vede nel file *mediated_schema.json***
- **Utilizzo di ChatGPT 4.0, fornendogli il file JSON, per definire gli attributi unificati a partire dagli attributi dei singoli file, come si vede nel file *merged_attribute.json***
- **Si definisce il file *mediated_schema.csv***
- **Riempimento manuale degli attributi per completare lo schema mediato.**

MEDIATED SCHEMA


```
"Name": [
    "./extracted_files/wissel-rappresentanti-ariregister.rik.ee.csv",
    "./extracted_files/wissel-aziende-info-clipper.com.csv",
    "./extracted_files/MalPatSaj-wikipedia.org.csv",
    "./extracted_files/MalPatSaj-forbes.com.csv",
    "./extracted_files/wissel-aziende-ariregister.rik.ee.csv",
    "./extracted_files/wissel-partners-ariregister.rik.ee.csv",
    "./extracted_files/AmbitionBox.csv"
],
"Trade Name": [
    "./extracted_files/wissel-aziende-info-clipper.com.csv"
],
"Address Name": [
    "./extracted_files/wissel-aziende-info-clipper.com.csv"
],
"Postalcode": [
    "./extracted_files/wissel-aziende-info-clipper.com.csv"
],
"City": [
    "./extracted_files/wissel-aziende-info-clipper.com.csv"
],
"State": [
    "./extracted_files/wissel-aziende-info-clipper.com.csv"
],
"Country": [
    "./extracted_files/MalPatSaj-forbes.com.csv",
    "./extracted_files/wissel-aziende-info-clipper.com.csv"
],
```


mediated_schema.json

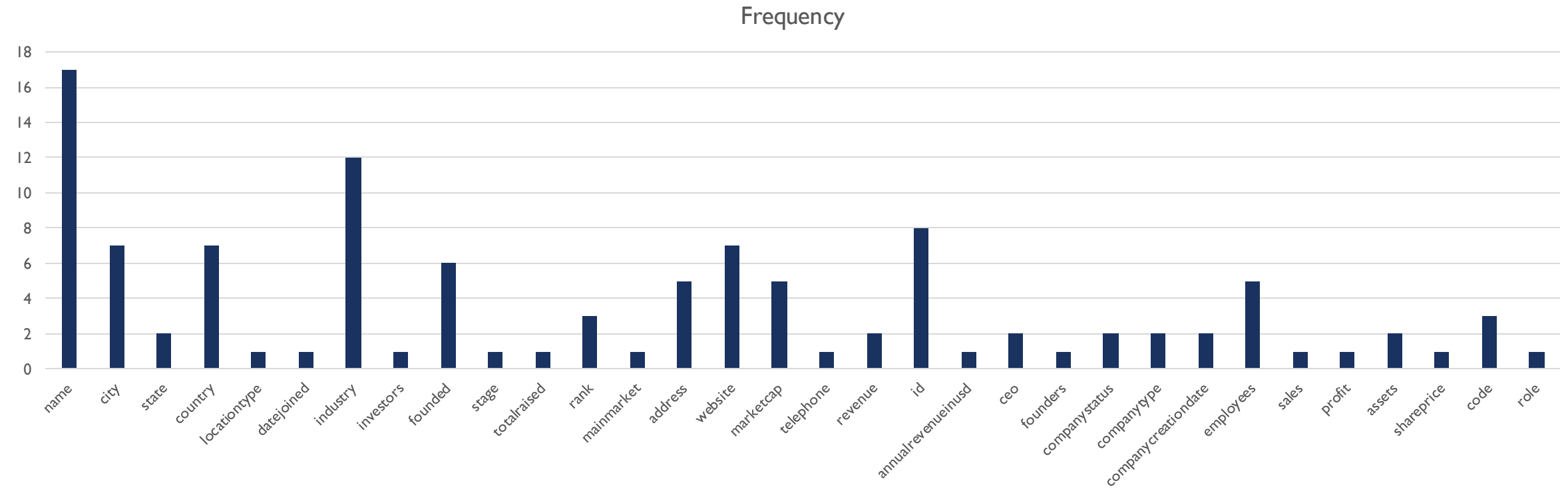

```
"name": {
    "unified_attribute": "name",
    "original_attributes": [
        "name",
        "Name"
    ],
    "files": [
        "./extracted_files/wissel-rappresentanti-ariregister.rik.ee.csv",
        "./extracted_files/DDD-teamblind.com.csv",
        "./extracted_files/MalPatSaj-forbes.com.csv",
        "./extracted_files/ft.com.csv",
        "./extracted_files/DDD-cbinsight.com.csv",
        "./extracted_files/output_govuk_bigsize.csv",
        "./extracted_files/output_globaldata.csv",
        "./extracted_files/valueToday_dataset.csv",
        "./extracted_files/wissel-partners-ariregister.rik.ee.csv",
        "./extracted_files/AmbitionBox.csv",
        "./extracted_files/hitHorizons_dataset.csv",
        "./extracted_files/wissel-aziende-info-clipper.com.csv",
        "./extracted_files/MalPatSaj-wikipedia.org.csv",
        "./extracted_files/disfold.com.csv",
        "./extracted_files/companiesMarketCap_dataset.csv",
        "./extracted_files/wissel-aziende-ariregister.rik.ee.csv"
    ]
},
"tradename": {
    "unified_attribute": "tradename",
    "original_attributes": [
        "Trade Name"
    ],
    "files": [
        "./extracted_files/wissel-aziende-info-clipper.com.csv"
    ]
},
"addressname": {
    "unified_attribute": "addressname",
    "original_attributes": [
        "Address Name"
    ],
    "files": [
        "./extracted_files/wissel-aziende-info-clipper.com.csv"
    ]
},
```


merged_attribute.json

MEDIATED SCHEMA

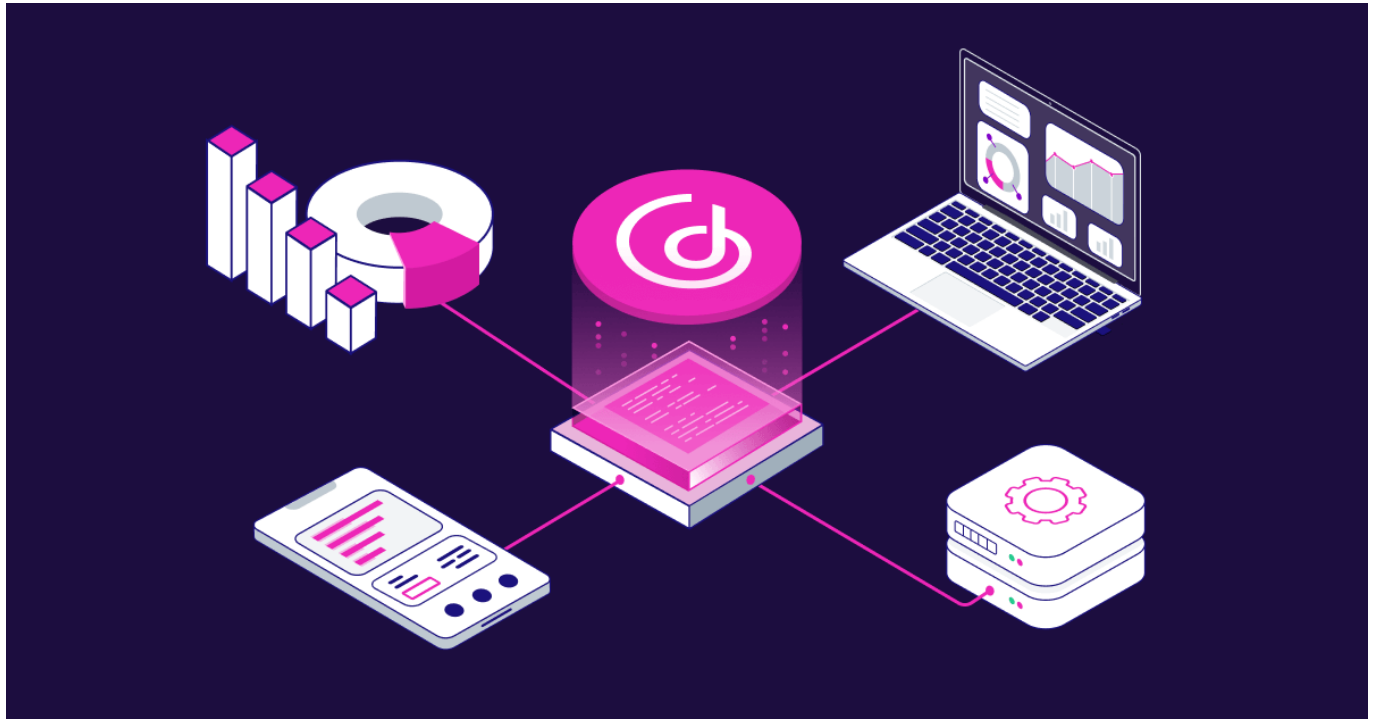
Unified Attribute	AmbitionBox.csv	DDD-cbinsight.com.csv	DDD-teamblind.com.csv	MalPatSaj-forbes.com.csv	MalPatSaj-wikipedia.org.csv	campaignindia.csv	companiesMarketCap_dataset.csv	company_social_urls.csv	disfold.com.csv	ft.com.csv
Name	Name	name	name	Name	Name	BRAND NAME	name	Company	name	name
City	headquarters	city	locations		Headquarters				headquarters	
Country		country		Country			country			country
Industry	Industry	industry	industry		Industry	CATEGORY	categories			industry
Founded	Foundation Year	founded	Founded		Founded					founded
Rank						RANK	rank			
investors		investors								
website			website						link	link
Market value				Market Value			market cap		market_cap	

MEDIATED SCHEMA - DISTRIBUTION



DATA INTEGRATION

Utilizzando i dataset di input e lo schema mediato viene effettuata l'operazione di data integration, trasformando e combinando i dati tra di loro, generando un singolo dataset che racchiude tutti i dati e tutti gli attributi dello schema mediato.



CONFRONTO CHE VALUTA LA SIMILARITÀ TRA LE COPPIE DI RECORD

RECORD LINKAGE

GROUND-TRUTH

Name 1	Name 2	Label
disney	walt disney company	1
apple	apple.inc	1
enel	enel energia spa	1
american express company	american express credit corporation	1
netflix global, llc	netflix worldwide entertainment, llc	1
sony music entertainment	sony pictures entertainment inc.	0
poste italiane	reti televisive italiane spa	0
ebay inc.	ebix inc.	0
new gold	new balance	0
airbus	airbnb	0

BLOCKING - PHONETIC

- **Generazione di un identificatore fonetico per ogni record (nome azienda).**
- **Genera un insieme di cluster sulla base dell'identificatore fonetico.**

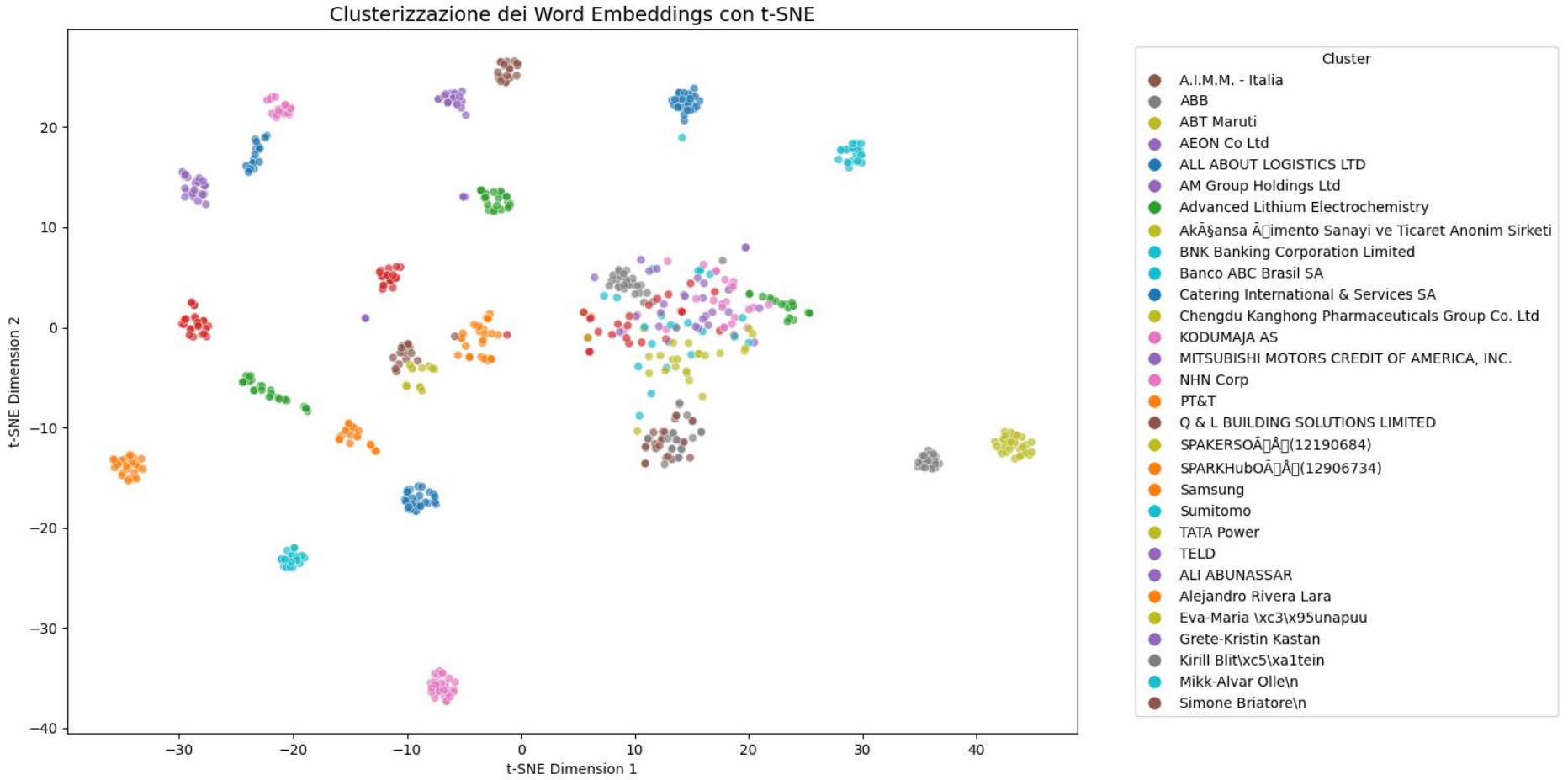
ID fonetico	Nomi associati
APLNKK	APPLE INC.; Ability, Inc.; Apple Inc
APLPRSSTMS	APOLLO POWER SYSTEMS; HBL Power Systems; Hbl Power Systems
PNKKPNKK	HP Inc; HP Inc.; eBay Inc; eBay Inc.

BLOCKING - EMBEDDINGS

- **Calcolo di rappresentazioni vettoriali (word-embeddings) basate sul modello 'Paraphrase-MiniLM-L6'.**
- **Genera un insieme di cluster composti da nomi di aziende considerate simili, usando DBSCAN come metodo di costruzione.**

ID Cluster	Nomi associati
924	AMAZON; AMAZON CORPORATE LLC; AMAZON DIGITAL UK LIMITED; AMAZON ONLINE UK LIMITED
1053	APPLE INC.; Apple; Apple Inc
10320	DISNEY INCORPORATED; Disney; THE WALT DISNEY COMPANY LIMITED; The Walt Disney Co

BLOCKING - EMBEDDINGS



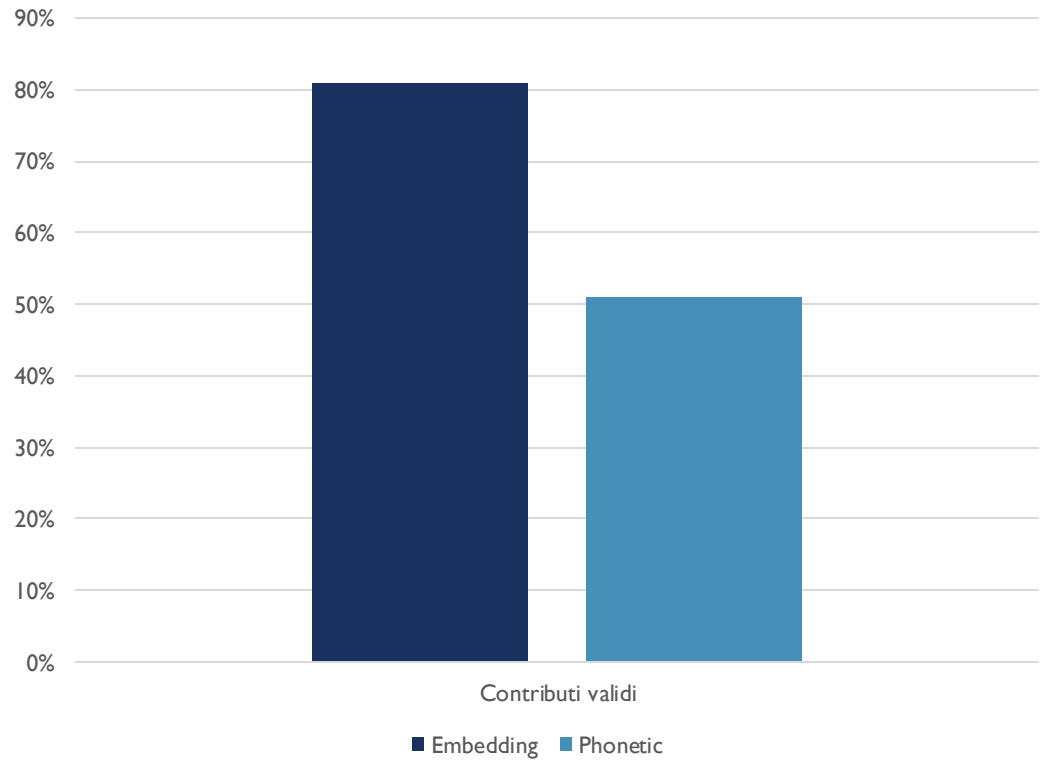
PAIRWISE MATCHING

- **Effettua un'iterazione su tutti i cluster calcolati nella fase di blocking.**
- **L'iterazione viene svolta sul rispettivo campo (ID fonetico o ID cluster) in base alla strategia utilizzata.**
- **Calcola la similarità tra tutte le possibili coppie di nomi, all'interno di ogni possibile cluster.**
- **Infine filtra solamente i risultati con un livello di similarità ≥ 0.5**

ID fonetico	AMSNLJSTKS	SNPKTRSNTRTNMNTNKK
Nome 1	amazon logistics	sony pictures entertainment inc.
Nome 2	yusen logistics	sony pictures home entertainment online inc.
Similarità	0.746	0.908

RISULTATI OTTENUTI - RECORDLINKAGE

- **Confronto tra i risultati ottenuti dai metodi di pairwise matching e la ground-truth.**
- **Si considera il contributo di una coppia valida solo se la coppia è presente nei risultati estratti e la similarità è superiore ad una soglia assegnata.**
- **Si considera il contributo di una coppia non valida solo se la coppia non è presente nei risultati estratti o la similarità è inferiore ad una soglia assegnata.**

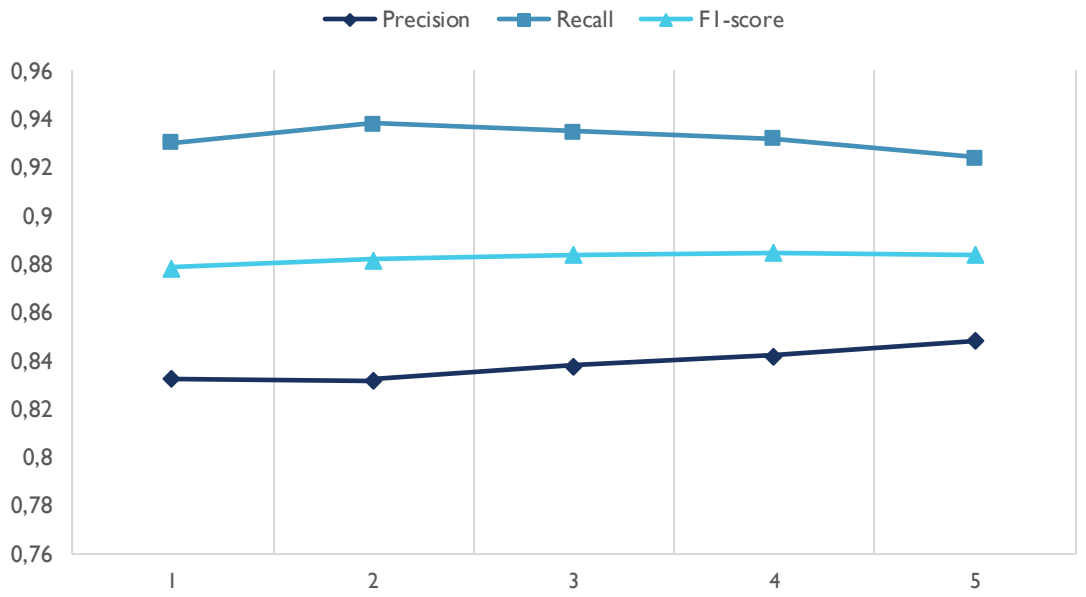


PAIRWISE MATCHING - DEEPMATCHER

Utilizzo del framework DeepMatcher basato su reti neurali, per identificare se due record rappresentano la stessa entità. Il modello è stato addestrato per 5 epoche su un dataset diviso in Train, Test e Validation set con la proporzione 60/20/20.

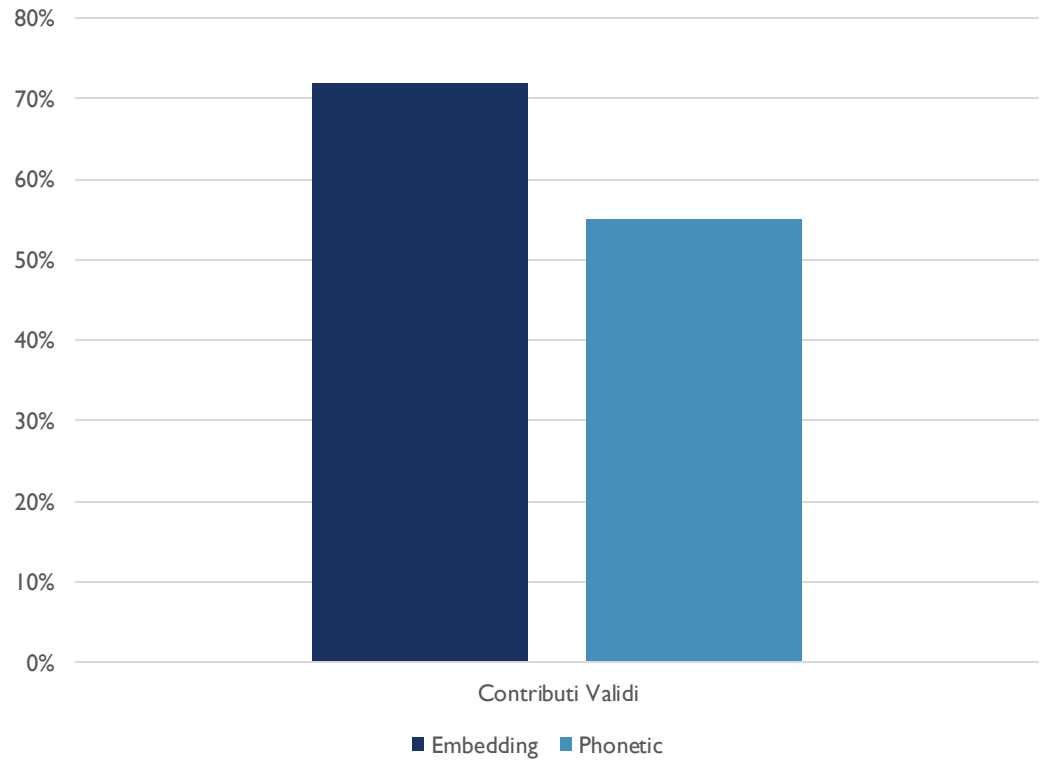
Il modello presenta la seguente accuratezza:

Precision	Recall	F-Measure
0,842	0,932	0,884



RISULTATI OTTENUTI - DEEPMATCHER

- **Confronto tra i risultati ottenuti dai metodi di pairwise matching e la ground-truth.**
- **Si considera il contributo di una coppia valida solo se la coppia è presente nei risultati estratti.**
- **Si considera il contributo di una coppia non valida solo se la coppia non è presente nei risultati estratti.**



PUNTI DI FORZA E LIMITAZIONI

- ☑ Efficienza strategie di Blocking
- ☑ Efficienza strategie di Pairwise Matching
- ☒ Metodo DeepMatcher richiede grandi quantità di dati

SVILUPPI FUTURI

Integrazione metodi ibridi per il Blocking

Valutazione su Dataset più ampi

Visualizzazioni interattive dei risultati

GRAZIE PER L'ATTENZIONE