Universitò degli studi Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione(DISCo)
Corso di laurea magistrale - Data Science

INTEGRATING KNOWLEDGE WITH NATURAL LANGUAGE PROCESSING FOR ENHANCED FACT-CHECKING

Relatore: Marco Viviani
Co-relatore: Rafael Penaloza Nyssen

Tesi di laurea magistrale
Simone Farallo
Matricola 889719

Introduzione

Contesto & Background

Disinformazione: diffusione intenzionale di informazioni false o fuorvianti con l'obiettivo di ingannare o manipolare le persone.
Fake News: informazioni false o fuorvianti che possono essere divulgate attraverso qualsiasi media allo scopo di produrre disinformazione.
Fact-checking: processo di verifica dell'accuratezza e della veridicità di affermazioni, dichiarazioni o informazioni.

Natural Language Processing (NLP)
Graph-based analytics

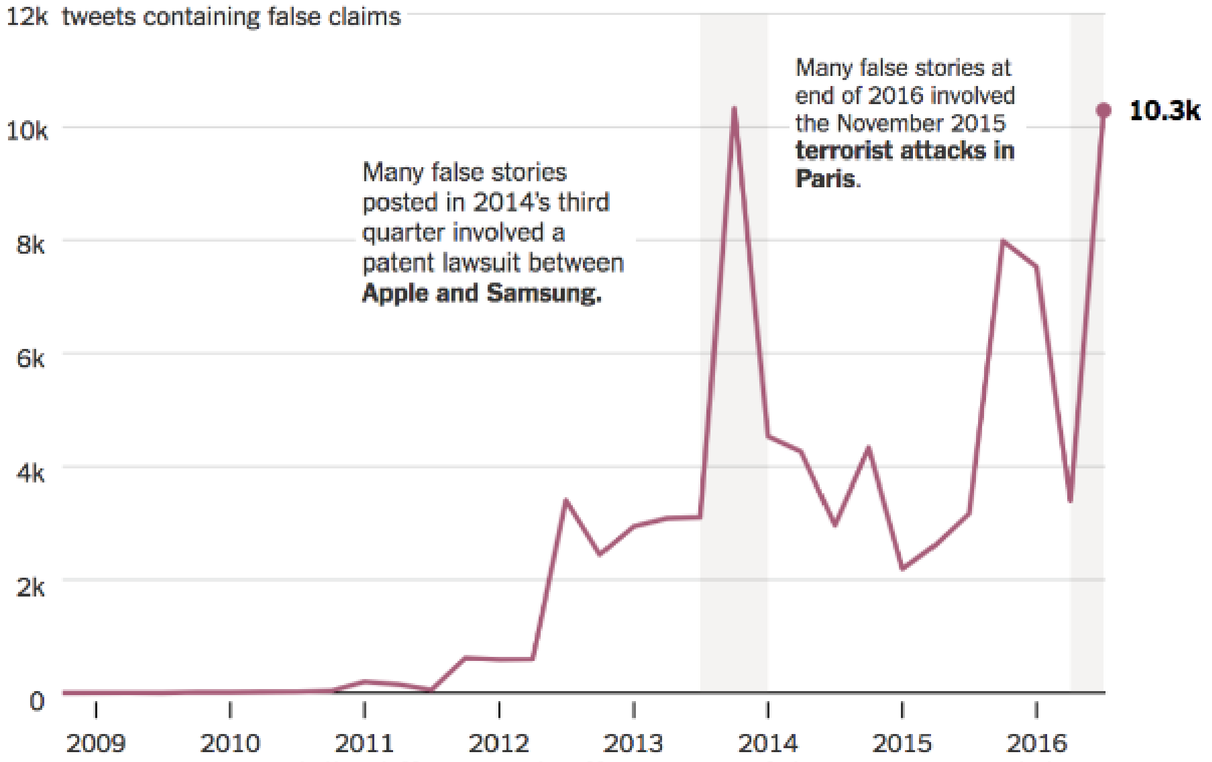


Figura: Frequenza della diffusione di affermazioni false su Twittter dal 2009 al 2016 - Fonte: New York Times

Introduzione

Obiettivi

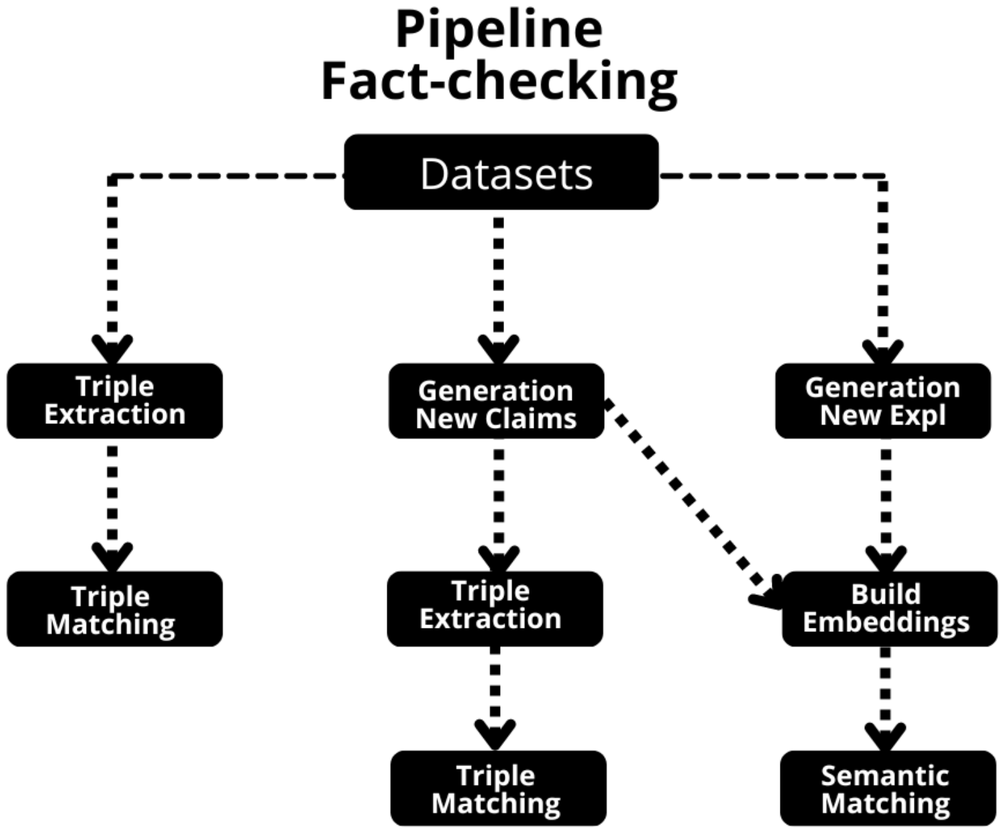
Analisi Comparativa: Valutazione delle diverse metodologie di fact-checking per evidenziarne punti di forza e limitazioni.

Utilizzo di **LLM** per il miglioramento dei sistemi di fact-checking attraverso l'introduzione di testo generato, per arricchire la semantica e il contenuto delle frasi.

Creazione di un ***Survey*** per valutare l'efficacia dei sistemi di fact-checking e promuovere l'avanzamento nel campo della disinformazione e delle fake-news.

Pipeline

1. Esplorazione dati
2. Costruzione della *Knowledge Base*
3. Estrazione Triple dalle affermazioni originali
4. Generazione di nuove affermazioni e spiegazioni
5. Estrazione Triple dalle nuove affermazioni
6. *Matching* Triple
7. *Matching* Semantico
8. Valutazione dei risultati



EDA & Preprocesing

Dataset Pubhealth (1)

(1). Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. 2018. doi: https://github.com/neemakot/Health-Fact-Checking.

Variabili d'interesse:

- claim
- explanation
- label
- subject (topic)

11,067 righe nei dataset:

- train: 9,832
- test: **1,235**

Costruzione **base di conoscenza** (*Knowledge base*) **:**

- **5,078** righe

claim_id	claim	date_published	explanation	fact_checkers	main_text	sources	label	subjects
2542	Study says too many Americans still drink too ...	February 25, 2013	On any given day in the United States, 18 perc...		That means the great majority of Americans sta...	http://bit.ly/X1NVtW	true	Health News
26678	Viral image Says 80% of novel coronavirus case...	March 13, 2020	The website Information is Beautiful published...	Paul Specht	Amid the spread of the novel coronavirus, many...	https://www.facebook.com/informationisbeautifu...	true	Facebook Fact-checks, Coronavirus, Viral image,

Figura: Prime 2 righe del dataset di test

EDA & Preprocesing

Esplorazione dei dati

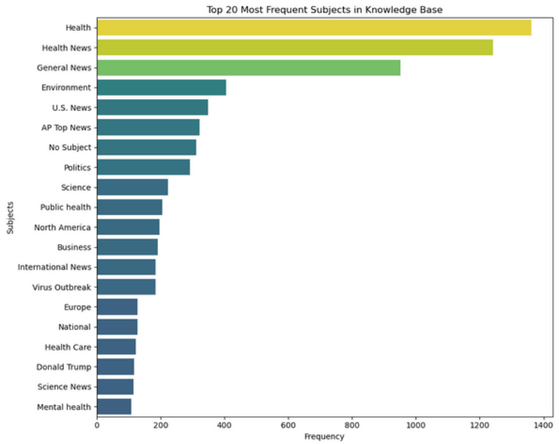


Figura: Topic più frequenti nella KB

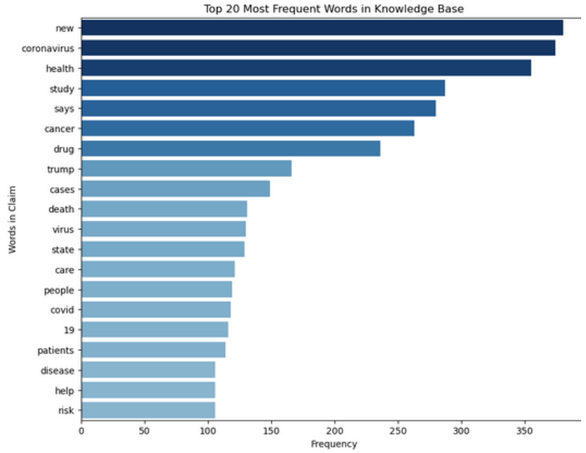


Figura: Parole più frequenti nella KB

Knowledge Base:

- 5078 affermazioni vere

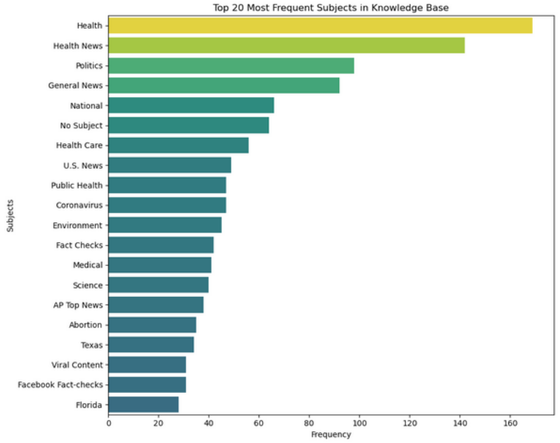


Figura: Topic più frequenti nel test

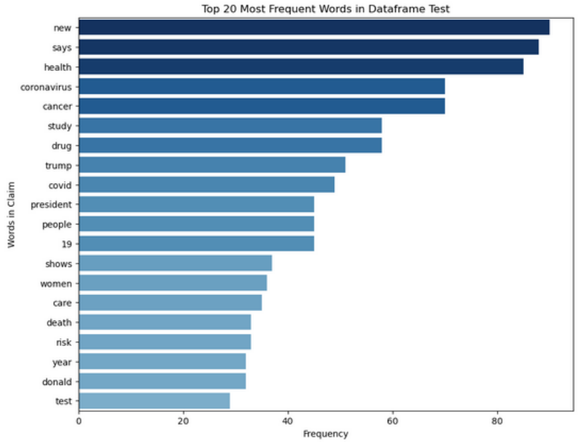


Figura: Parole più frequenti nel test

Test:

- 599 affermazioni “vere”
- 636 alltre affermazioni

Estrazione triple e generazione testo

Metodologie

Triple SPO: strutture dati che rappresentano informazioni o relazioni in forma di affermazioni, dove il soggetto è collegato all'oggetto tramite un predicato.

Metodologie e modelli utilizzati:

- **REBEL** (*Autoregressive seq2seq models*)
- **CoreNLP** (*Open information extraction-OpenIE*)

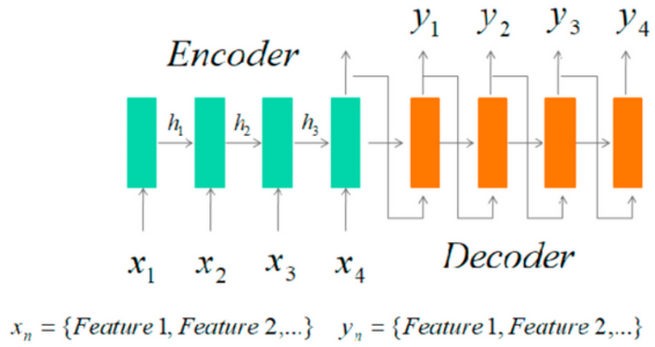


Figura: Autoregressive seq2seq models

Generazione testo: generazione di nuove frasi e nuove spiegazioni.

Metodologie e modelli utilizzati:

- **GPT-3.5**
- Prompt engineering

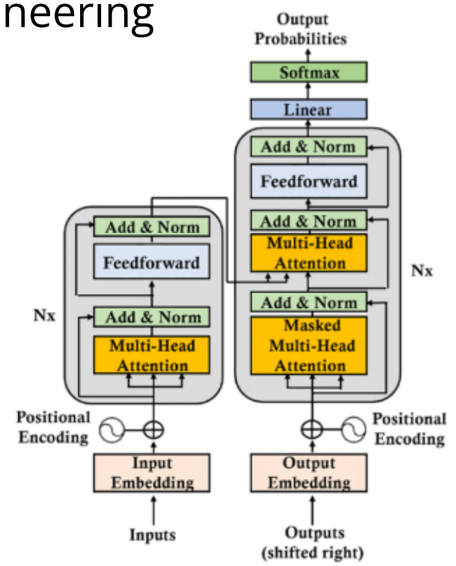


Figura: Modelli Transforrmers (GPT)

Estrazione triple e generazione testo

Esplorazione risultati

Estrazione triple effettuata su affermazioni originali e generate, sia sulla base di conoscenza che sul test.

Modelli	Tot. Triple	Soggetti un.	Predicati un.	Ogetti un.
REBEL	16977	8398	187	5990
CoreNLP	16291	4691	3994	10628

Tabella: Triple estratte dalle affermazioni originale della KB

Modelli	Tot. Triple	Soggetti un.	Predicati un.	Ogetti un.
REBEL	17,354	7,897	188	5,624
CoreNLP	33,438	5,937	5,161	17,320

Tabella: Triple estratte dalle affermazioni generate della KB

Generazione nuove affermazioni e nuove spiegazioni.

claim	new_claim
Opossums kill thousands of ticks each week, inhibiting the spread of Lyme Disease to humans.	"Opossums play a crucial role in controlling the spread of Lyme Disease by consuming thousands of ticks weekly, thus reducing the risk of transmission to humans."
Democrats hoping to flip House not just trash-talking Trump.	"The Democrats are focused on flipping the House, not just criticizing Trump, who is a member of the Democratic party they oppose."

Figura: Affermazioni originali e generate

explanation	new_expl
What's true: Some data indicate opossums eat thousands of deer ticks per season, reducing the number that can go on to spread Lyme Disease to humans. What's false: How much of an impact opossums' eating ticks has on Lyme Disease infection rates is indeterminate.	Opossums play a crucial role in reducing the spread of Lyme Disease by consuming thousands of ticks per season. However, the exact impact of opossums on Lyme Disease infection rates remains uncertain.
Democrats hoping to flip enough seats to regain control of the U.S. House of Representatives say they aren't putting all their eggs in the anti-Trump basket.	Democrats aiming to win back control of the U.S. House of Representatives are not solely relying on criticizing Trump.

Figura: Spiegazioni originali e generate

Matching

Metodologie e logiche

Assunzione di **modello chiuso**: le affermazioni assenti nella base di conoscenza sono considerate false.

Matching Triple: Costruzione di ***grafi*** e verifica delle affermazioni utilizzando triple estratte e algoritmi.

Logiche di Valutazione:

- **Metodo A**: se almeno una tripla di un dato testo X, corrisponde con successo a una qualsiasi tripla della base di conoscenza, il testo X viene etichettato come vero
- **Metodo B**: il testo X è considerato vero se e solo se più del 50% delle sue triple trovano una corrispondenza nella base di conoscenza.

Matching Semantico: Costruzione di ***embeddings*** utilizzando il testo generato e verifica delle affermazioni tramite la ***cosine similarity***.

Logiche di Valutazione:

- Un testo X è considerato vero se è simile almeno alla '*soglia*' di uno dei testi nella base di conoscenza.
- S*oglie*: 50%, 70%, 90%

Matching Triple

Implementazione

Costruzione **grafi** con **triple SPO**

Grafo	Nodi	Archi
REBEL (Aff. originali)	11,334	10,854
CoreNLP (Aff. originali)	14,318	15,223
REBEL (Aff. generate)	10,468	10,618
CoreNLP (Aff. generate)	21,643	29,200

Tabella: Nodi e archi dei 4 grafi.

Algoritmi:

- **Fuzzy** matching
- **ShortestPath** con discriminante
- Node Embedding (**Nod2Vec**) con discriminante
- **Knowledge Linker** con discriminante

Discriminante sul predicato:

- Fuzzy
- BERT

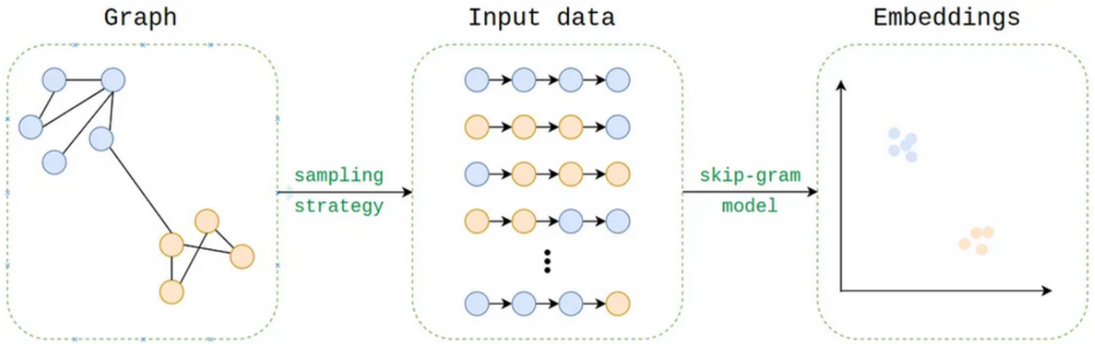


*Tabella: Node2vec: processo costruzione **embedding***

Matching Semantico

Implementazione

Embeddings su nuove affemraizoni e spiegazioni

- Adaptive Discriminator Augmentation (**ADA**)
- Robustly optimized BERT approach (**RoBERTa**)
- Enhanced Representation through knowledge Integration (**ERNIE**)
- **DistilBERT**

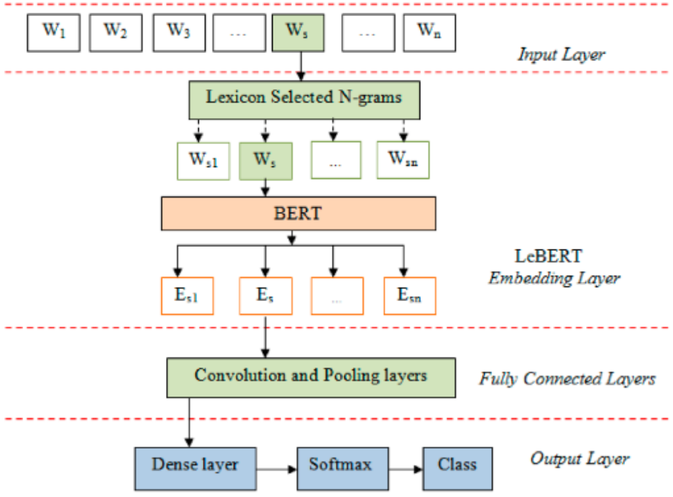


Figura:BERT

Ricerca di similiarità:

- Facebook AI Similarity Search(**FAISS**)

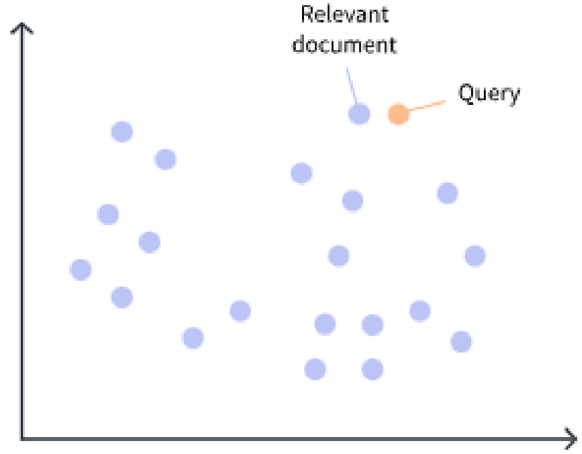


Figura:Match semantico con FAISS

Risultati

Matching Triple

Metodo	Triple	Algoritmo	*Accuracy*	F1-*score*
A	REBEL	ShortPath_bert	0.578	0.480
A	CoreNLP	KL_bert	0.531	0.318
B	REBEL	KL_bert	0.570	0.450
B	CoreNLP	ShortPath_bert	0.588	0.460

Tabella: Risultati migliori ottenuti per ogni metodo e tipo di tripla sulle affermazioni originali

Metodo	Triple	Algoritmo	*Accuracy*	F1-*score*
A	REBEL	ShortPath_bert	0.600	0.510
A	CoreNLP	Nod2vec_bert	0.511	0.407
B	REBEL	Nod2vec_bert	0.540	0.400
B	CoreNLP	ShortPath_bert	0.628	0.470

Tabella: Risultati migliori ottenuti per ogni metodo e tipo di tripla sulle affermazioni generate

Risultati

Matching Triple

Valutazione risultati su **affermazioni originali**:

Metodo A:

- Risultati migliori utilizzando le triple REBEL rispetto a CoreNLP.
- Performance migliore: ***ShortPath_bert*** con REBEL.

Metodo B:

- Risultati simili tra triple REBEL e CoreNLP.
- Performance migliore: ***ShortPath_bert*** e ***Node2Vec_bert*** con CoreNLP.
- ***KL_fuzzy*** con CoreNLP: Ottima precisione, ma recall basso.

Valutazione risultati su **affermazioni generate**:

Metodo A:

- Risultati omogenei le triple REBEL rispetto a CoreNLP.
- Risultati leggermente migliori alle affermazioni originali.
- Performance migliore: ***ShortPath_bert*** e ***Node2Vec_bert*** con REBEL.

Metodo B:

- Risultati migliori utilizzando CoreNLP.
- Risultati migliori rispetto alle affermazioni originali.
- Performance migliore: ***Node2Vec_bert*** e ***KL_bert*** con CoreNLP.

Risultati

Matching Semantico

Soglia	Modello	*Accuracy*	*F1-score*
0.5	ADA	0.610	0.633
0.7	ADA	0.577	0.210
0.9	ERNIE	0.520	0.647

Tabella: Risultati migliori ottenuti per ogni soglia sulle affermazioni generate

Soglia	Modello	*Accuracy*	*F1-score*
0.5	ADA	0.630	0.657
0.7	ADA	0.577	0.232
0.9	DistilBERT	0.555	0.640

Tabella: Risultati migliori ottenuti per ogni soglia sui riassunti generati

Risultati

Matching Semantico

Valutazione risultati su **affermazioni generate**:

- ADA: performance notevoli a una soglia di somiglianza di 0.5. Tuttavia, l'incremento della soglia a 0.7 e 0.9 mostra un significativo calo delle performance.
- RoBERTa ed ERNIE: Mostrano performance consistenti a tutte le soglie.
- DistilBERT: a una soglia di 0.9 riesce a mantenere un precisione alta e un F1-score migliorato rispetto ad ADA.
- **Miglior modello**: ADA a una soglia di somiglianza di 0.5.

Valutazione risultati su **riassunti generati:**

- ADA: Miglior equilibrio tra precisione e recall a una soglia di 0.5, rendendolo adatto per la verifica di spiegazioni rilevanti.
- **distilBERT**: Consistenza notevole nella recall a soglie di 0.5 e 0.7.
- RoBERTa e ERNIE hanno limiti di computazione non indifferenti.
- **Miglior modello**: distilBERT a soglia di 0.9.

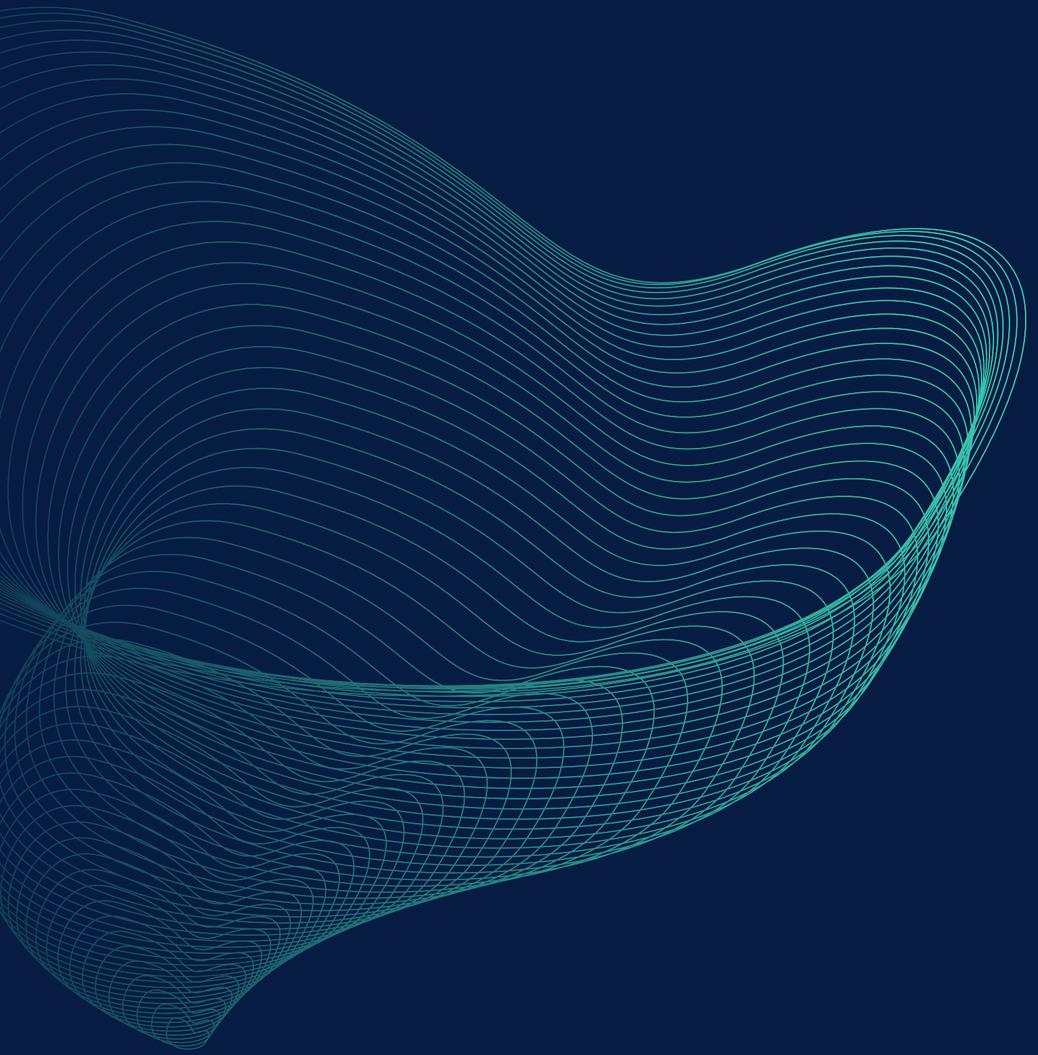
Conclusioni e lavori futuri

L'analisi comparativa tra diversi metodi di estrazione delle triple e l'uso di modelli generativi per la creazione di nuovi testi ha sottolineato il potenziale dell'integrazione delle tecniche tradizionali di NLP con le capacità avanzate degli LLM:

- le tecniche basate sui grafi hanno avuto un aumento delle performance con le frasi generate; ma queste inseriscono un ulteriore strato di **casualità** nel sistema.
- il **matching semantico** è risultato un metodo più efficace, ma più costoso dal punto di vista computazionale.
- l'utilizzo di algoritmi su grafi con struttura semplice hanno evidenziato la **bassa capacità** di questi modelli nel catturare la **semantica**.

Lavori futuri:

- **Ampliamento della base di conoscenza**
- **Analisi semantica approfondita**
- **Integrazione di grafi di conoscenza con ontologie**
- **Aumento potenza computazionale**



Grazie per l'attenzione