UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

RELAZIONE DEL PROGETTO D'ESAME

DEEP LEARNING

Anonimizzazione e Text Summarization di Dati Medici

Studente:

Giovanni Borrelli
0522501807

ANNO ACCADEMICO 2023/2024

Sommario

1 INTRODUZIONE

1.1 Introduzione al problema e all'obiettivo del progetto

Lo scopo di questo progetto è esplorare tecniche di anonimizzazione automatica di dati clinici tramite modelli di deep learning, con particolare attenzione all'utilizzo di BERT, adattato al dominio biomedico, per Named Entity Recognition (NER). L'obiettivo è rimuovere o mascherare in modo efficace le informazioni sensibili all'interno di testi medici. Successivamente, si applica una fase di text summarization, sfruttando modelli linguistici avanzati come ChatGPT, guidati da prompt progettati ad hoc e arricchiti con le etichette ottenute dal NER. Il flusso di lavoro proposto mira non solo a garantire la privacy del paziente, ma anche a preservare la significatività clinica dei testi originali. L'interesse principale del progetto è valutare quanto la summarization, eseguita dopo l'anonimizzazione, riesca a mantenere il significato del contenuto originario senza reintrodurre elementi sensibili.

1.2 Research questions individuate

RQ1: Quali tecniche basate su modelli di deep learning sono più efficaci per l'anonimizzazione automatica di dati clinici testuali, e in che modo BERT con Named Entity Recognition può essere utilizzato a questo scopo?

RQ2: In che modo è possibile applicare tecniche di text summarization, basate su prompt engineering e arricchite con le informazioni semantiche estratte tramite NER, per generare riassunti anonimizzati che mantengano la significatività clinica dei testi originali?

2 BACKGROUND E STATO DELL'ARTE

La protezione della privacy nei testi clinici è un tema centrale nella ricerca biomedica, specialmente in un contesto in cui l'uso di dati testuali per l'addestramento di modelli di deep learning è in forte crescita. Le tecniche di anonimizzazione automatica si sono evolute per garantire la rimozione o mascheramento di informazioni personali identificabili (PII), pur mantenendo l'utilità clinica dei testi. Tra gli approcci più diffusi si trovano:

- metodi **rule-based**, che sfruttano pattern regolari per identificare nomi, date, luoghi (es. regex e dizionari),
- metodi **statistici**, che si basano su modelli sequenziali come CRF,
- e metodi **deep learning-based**, che utilizzano reti neurali per il riconoscimento contestuale delle entità sensibili.

In particolare, i modelli di Named Entity Recognition (NER) addestrati su dati clinici si sono rivelati efficaci per questa finalità. Il modello BERT, ottimizzato per compiti di NER biomedico, ha ottenuto risultati all'avanguardia nel riconoscimento di entità cliniche come sintomi, diagnosi, farmaci, strutture biologiche e procedure terapeutiche, dimostrandosi particolarmente adatto per compiti di anonimizzazione nei documenti sanitari. https://arxiv.org/abs/1901.08746

Un contributo recente in questa direzione è quello di Gounley et al. (2024), che propongono un approccio ibrido basato su modelli NER e regole di sostituzione semantica, applicato a cartelle cliniche reali. La pipeline automatica da loro proposta anonimizza entità sensibili come nomi, date e località mantenendo la leggibilità e l'informatività del testo. Questo lavoro mostra che combinare NER e tecniche di mascheramento semantico è un modo efficace per proteggere i dati clinici, senza renderli inutili per le applicazioni successive.
Ad esempio, una tecnica di mascheramento consiste nel sostituire l'età del paziente con un placeholder generico come **[age]**, in modo da nascondere l'informazione sensibile mantenendo però la struttura e il significato del testo. https://arxiv.org/pdf/2412.08255

Un secondo tema centrale riguarda la possibilità di addestrare modelli NLP su dati già anonimizzati. Diverse ricerche hanno dimostrato che, se l'anonimizzazione è eseguita in modo conservativo (cioè preservando le strutture linguistiche e le entità cliniche non sensibili), è possibile mantenere elevate prestazioni nei task NLP come classificazione, estrazione di relazioni o question answering. Alcuni studi (Neamatullah et al., 2008) mostrano che modelli addestrati su testi de-identificati mantengono la capacità di generalizzare, a patto che l'anonimizzazione non rimuova o alteri contenuti clinici rilevanti. In questo contesto, i modelli come BioBERT, già pre-addestrati su grandi corpora biomedici, possono essere riutilizzati o fine-tuned su dati anonimizzati con minime perdite di accuratezza. https://doi.org/10.1186/1472-6947-8-32

Infine, la text summarization rappresenta uno strumento strategico non solo per la compressione dell'informazione, ma anche per la generazione di testi clinici sintetici, controllati e anonimizzati. L'utilizzo di modelli generativi come GPT-3 o GPT-4 permette di creare riassunti coerenti e privi di dati sensibili, soprattutto se guidati tramite prompt engineering basati sulle etichette NER precedentemente estratte. In questo modo è possibile condizionare la generazione in modo da enfatizzare (o escludere) determinati concetti clinici, garantendo un equilibrio tra privacy e mantenimento della significatività medica. Studi recenti (Zhang et al., 2023; Moramarco et al., 2023) confermano che i LLM, se correttamente guidati, riescono a produrre riassunti informativi che riducono la probabilità di re-identificazione, ponendosi come una soluzione promettente in contesti sanitari. https://arxiv.org/abs/2303.11032 e https://arxiv.org/abs/2412.12040

2.1 Analisi dei principali approcci esistenti

Nel campo dell'anonimizzazione dei dati clinici, come accennato precedentemente, i principali metodi attualmente utilizzati si dividono in tre categorie: approcci rule-based, modelli statistici e modelli deep learning. I metodi rule-based, pur essendo semplici da implementare, spesso soffrono di scarsa adattabilità e non sono robusti a variazioni linguistiche o contesti diversi. I modelli statistici come CRF migliorano la flessibilità ma possono risultare limitati quando si tratta di catturare informazioni più complesse o contestuali.

I modelli deep learning, in particolare quelli basati su architetture Transformer come BioBERT, rappresentano lo stato dell'arte per il riconoscimento delle entità sensibili grazie alla loro capacità di comprendere il contesto clinico. Tuttavia, la sola identificazione delle entità non risolve completamente il problema dell'anonimizzazione: è necessaria una fase di mascheramento o sostituzione che preservi la coerenza e la validità clinica del testo.

Un problema comune nei metodi attuali è che non esiste ancora un modo standard per bilanciare la protezione della privacy con il mantenimento dell'utilità dei dati anonimizzati, soprattutto quando questi dati vengono poi usati per addestrare modelli di NLP. Inoltre, spesso manca una valutazione chiara di come le tecniche di anonimizzazione influenzino le prestazioni dei modelli che utilizzano quei dati.

Per questo motivo, è importante sviluppare soluzioni che uniscano il riconoscimento automatico delle informazioni sensibili con tecniche di mascheramento mirate, accompagnate da metodi di valutazione che misurino sia la sicurezza sia l'efficacia dell'anonimizzazione. Questo progetto vuole proprio affrontare tali sfide, cercando di proteggere i dati clinici senza perdere la qualità delle analisi successive.

3 METODOLOGIA

3.1 Descrizione del dataset

Il dataset utilizzato per questo progetto è disponibile pubblicamente su Kaggle al seguente link: https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions.

Il dataset contiene circa **5.000 trascrizioni mediche** reali in lingua inglese. Ogni trascrizione è associata a una **specializzazione medica**, come *chirurgia*, *cardiologia*, *pediatria* e altre. Le due colonne più rilevanti per l'analisi sono:

- la trascrizione testuale del referto medico;
- la specializzazione clinica a cui esso appartiene.

Queste informazioni permettono di valutare l'efficacia dei modelli di anonimizzazione e summarization in contesti clinici eterogenei e specifici.

3.1.1 Data Cleaning

Prima di procedere con le fasi di anonimizzazione e sintesi del testo, sono state effettuate alcune operazioni di pulizia dei dati per garantire la qualità del dataset. In particolare, sono stati **rimossi i valori nulli**, ossia le trascrizioni mancanti o incomplete; **eliminati i duplicati**, ovvero trascrizioni identiche presenti più volte nel dataset; e **filtrati i valori anomali**, come testi troppo brevi o privi di contenuto medico rilevante.
Queste operazioni hanno permesso di ottenere un insieme di dati più coerente, riducendo il rumore e migliorando l'affidabilità delle analisi successive.

3.2 NER

Il modello BERT utilizzato è disponibile al seguente link:

https://huggingface.co/d4data/biomedical-ner-all

Per quanto riguarda le label utilizzate per il NER, sono 83, e si suddividono in 7 categorie:

1. Informazioni sul paziente

Riguardano dati anagrafici, storico personale, stile di vita o informazioni familiari che descrivono il profilo del paziente.

2. Concetti clinici e biomedici

Includono termini relativi a malattie, sintomi, strutture anatomiche e altri concetti fondamentali del linguaggio medico e biologico.

3. Diagnostica e trattamenti

Comprendono procedure diagnostiche o terapeutiche, farmaci, somministrazioni e valori di laboratorio associati alla diagnosi e alla cura.

4. Quantità, tempo e misure

Rappresentano entità misurabili o temporali come durata, frequenza, grandezze fisiche e riferimenti temporali.

5. Attributi descrittivi

Descrivono caratteristiche qualitative o visive associate ad altre entità, come aspetto, forma, consistenza o concetti soggettivi.

6. Luoghi e localizzazioni

Riguardano spazi o posizioni fisiche non biologiche associati al contesto clinico, come ambienti o strutture.

7. Altro

Categoria rimanente che include entità generiche, eventi non specificamente clinici, riferimenti testuali e soggetti menzionati.

Durante l'aggregazione delle predizioni delle parole in etichette, ho preferito usare **aggregation_strategy="first"** invece di "**simple**" perché quest'ultima tendeva a suddividere frequentemente una parola in più token e assegnare etichette separate a ciascun token. Questo causava frammentazione e incoerenza nelle entità riconosciute.
Al contrario, la strategia **first** assegna l'etichetta del primo token a tutta la parola (composta da più sub-token), garantendo una maggiore coerenza e stabilità nell'identificazione delle entità, con un risultato complessivamente più preciso e utilizzabile.

3.2.1 Anonimizzazione

A questo punto, dopo aver identificato le etichette corrispondenti a dati medici sensibili, è stata applicata una tecnica di mascheramento, sostituendo ogni parola o sequenza di parole riconosciute con la rispettiva etichetta.

Ad esempio, la frase *"Il paziente ha una diagnosi di diabete"* viene trasformata in *"Il paziente ha una diagnosi di [Disease_disorder]"*.

3.3 Text Summarization

La Per la fase di **text summarization** è stato utilizzato il modello **GPT-4o** applicato su un sottoinsieme del dataset di 5.000 trascritti medici, già opportunamente anonimizzati tramite riconoscimento delle entità sensibili e mascheramento semantico. Ogni trascrizione è stata elaborata manualmente in ChatGPT, chiedendo la generazione di un riassunto informativo e conciso del testo clinico, facendo riferimento anche a quali sono state le label più frequenti identificate in fase di analisi, mantenendo le informazioni clinicamente rilevanti ma senza reinserire dettagli identificabili.

Per valutare la qualità dei riassunti generati, è stato adottato un doppio criterio:

1. **Qualità semantica del riassunto**:
   È stata misurata tramite **BERTScore**, una metrica basata su modelli pre-addestrati che confronta i vettori di embedding del riassunto e del testo originale. In questo modo valuta la similarità semantica tra il trascritto e il riassunto generato.

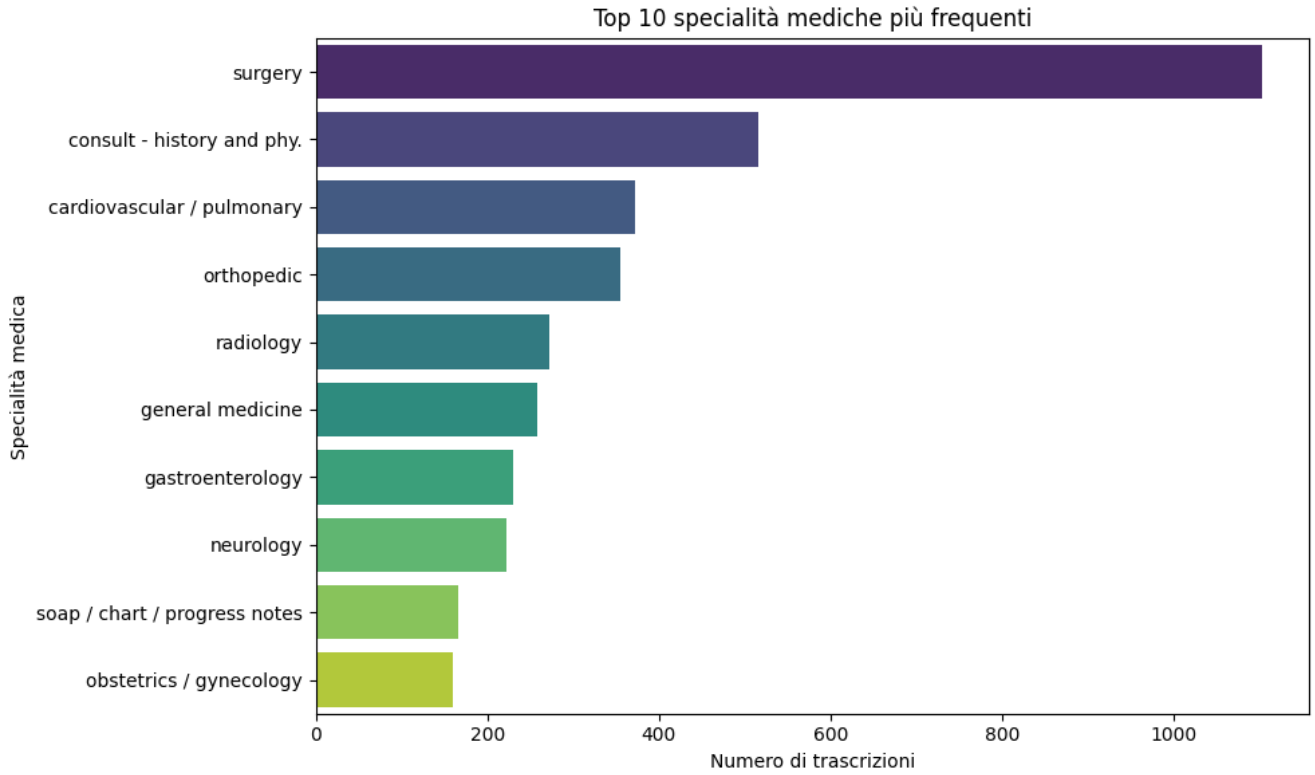
2. **Verifica dell'anonimizzazione nel riassunto**:
   È stato contato il numero di etichette NER presenti nel riassunto e confrontato con quelle del testo originale. Questo ha permesso di capire se il riassunto conservasse un numero adeguato di concetti clinici **senza reinserire entità sensibili**.
   Idealmente, il riassunto dovrebbe **ridurre il numero di label irrilevanti o potenzialmente identificabili** (come nomi, età, luoghi) preservando invece le etichette cliniche principali (es. *Sign_symptom*, *Disease_disorder*, *Therapeutic_procedure*).
   Quindi, non è necessariamente meglio avere lo stesso numero di label, ma è preferibile che il riassunto contenga meno etichette, a condizione che siano rappresentative delle informazioni mediche principali.

4 VALUTAZIONE SPERIMENTALE

4.1 Caratteristiche del Dataset

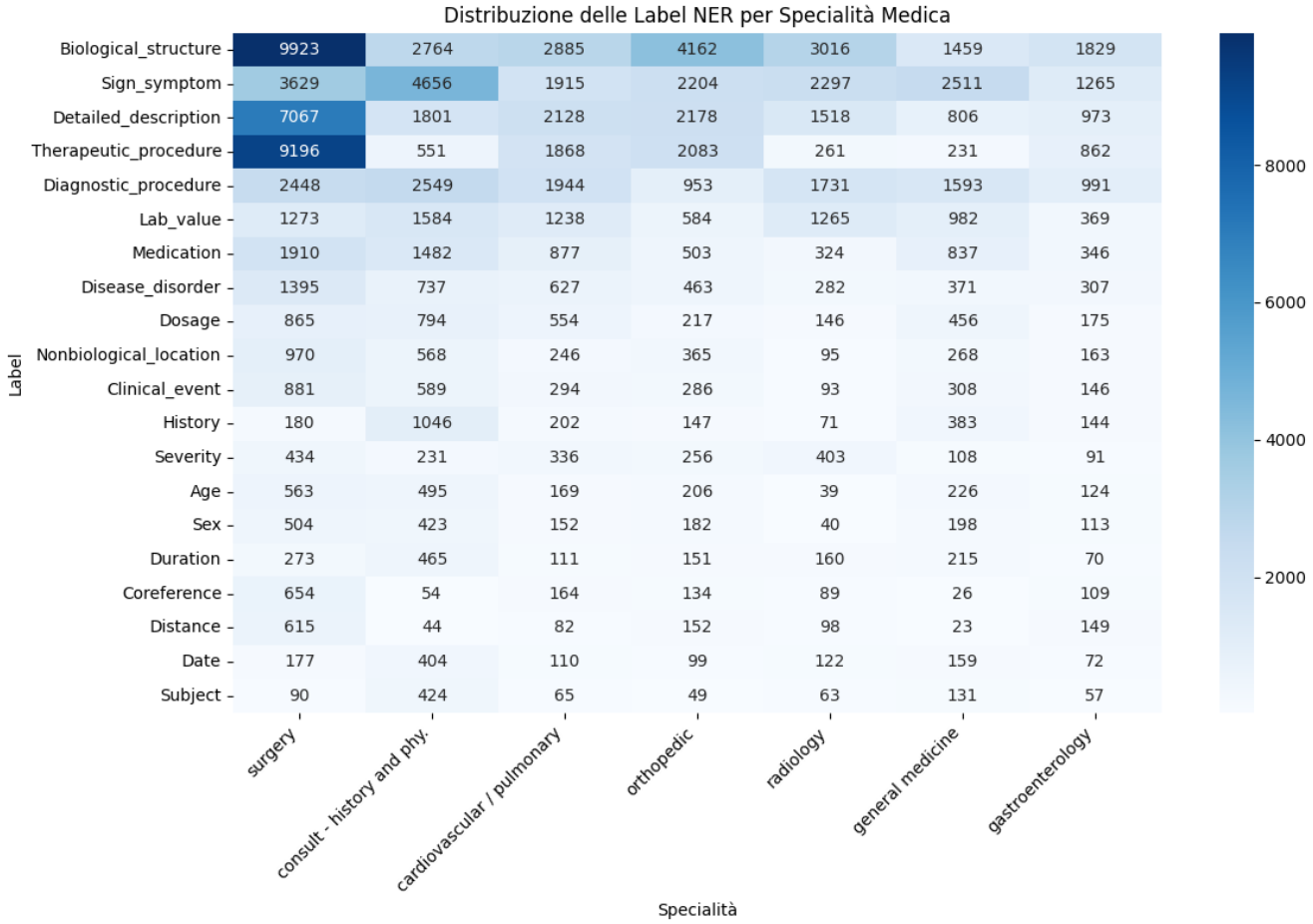
Gli attributi più significativi del dataset sono la “specializzazione medica” (che può essere, ad esempio, chirurgia o cardiologia) e il “trascritto”, che rappresenta il contenuto vero e proprio del trascritto medico. Esiste anche una terza colonna chiamata “keywords”, ma dopo averne analizzato il contenuto, è quasi uguale alla colonna “specializzazione medica”, in quanto quasi sempre la keyword coincide proprio con il tipo di specializzazione.

Ecco un grafico che mostra la frequenza delle prime 10 specializzazioni mediche:



La specializzazione *surgery* rappresenta oltre il 20% dei trascritti (1103 su 5000), indicando che il dataset è sbilanciato verso ambiti chirurgici.
Oltre alla chirurgia, sono ben rappresentate anche aree come *cardiologia/pneumologia*, *ortopedia*, *radiologia*, *neurologia*, *gastroenterologia* e *medicina generale*, suggerendo una buona copertura di specializzazioni cliniche diverse, utili per analisi eterogenee. La composizione dei trascritti mostra una distribuzione non perfettamente bilanciata ma sufficientemente varia.

A questo punto, si vuole analizzare se la specializzazione medica influisce sulle label identificate da BERT. Di seguito è possibile osservare una heatmap che mostra, per ciascuna delle 10 specialità mediche più ricorrenti, la frequenza di occorrenza delle diverse etichette NER. È uno strumento utile per valutare se e come la distribuzione semantica delle entità varia tra specializzazioni cliniche.



Alcune label sono molto frequenti in tutte le specialità:

Biological_structure, Detailed_description, Therapeutic_procedure, Sign_symptom sono le più frequenti trasversalmente.

Ad esempio, Biological_structure domina in surgery (9923), ma è alta anche in orthopedic (4162), radiology (3016) e persino in specialità meno invasive.

Ci sono entità comuni a quasi tutte le discipline cliniche, riflettendo un linguaggio medico condiviso.

In consult - history and phy., si osserva un uso relativamente alto di Sign_symptom e History, coerente con la natura di visita anamnestica e raccolta sintomi.

In cardiovascular / pulmonary compaiono con frequenze medie Lab_value, Disease_disorder, Dosage, coerenti con patologie croniche trattate in quella specializzazione.

Quindi esistono lievi variazioni tematiche, ma non così marcate da identificare specialità solo dalla distribuzione delle label.

Ciò suggerisce che, dal punto di vista semantico, **i trascritti medici utilizzano un linguaggio comune, indipendentemente dalla specializzazione.**

4.2 Analisi dei risultati del NER

È stato effettuato un confronto tra le etichette trovate dal modello NER e quelle annotate manualmente su un sottoinsieme dei trascritti. Come metriche di valutazione sono state utilizzate:

Precision (P): la precision è la frazione di previsioni corrette tra tutte le previsioni fatte. Si calcola come il numero di veri positivi (True Positives, TP) diviso per la somma dei veri positivi e dei falsi positivi (False Positives, FP):

$$P = \frac{TP}{TP + FP}$$

Recall (R): il recall è la frazione di oggetti correttamente rilevati tra tutti gli oggetti presenti. Si calcola come il numero di veri positivi diviso per la somma dei veri positivi e dei falsi negativi (False Negatives, FN):

$$R = \frac{TP}{TP + FN}$$

F1 score (F1): l'F1 score è la media armonica tra precision e recall, fornendo una misura bilanciata della accuratezza del modello. Si calcola come:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Le metriche di valutazione ottenute sono state approssimativamente:

- Precision: 0.81
- Recall: 0.88
- F1 score: 0.84

Questi valori indicano che spesso il modello rileva anche entità che corrispondono a dati non sensibili o meno rilevanti. Questo risultato è comunque positivo, perché è preferibile intercettare più entità, anche se non tutte sensibili, piuttosto che rischiare di non rilevare dati sensibili importanti. In altre parole, meglio un riconoscimento non importante che una mancanza, soprattutto in ambito di tutela della privacy.

4.2.1 RQ1

L'**anonimizzazione automatica di dati clinici testuali** rappresenta una sfida cruciale per la tutela della privacy, soprattutto considerando la presenza di molteplici tipologie di informazioni sensibili (nomi, date, luoghi, numeri di identificazione, ecc.). Ci siamo quindi chiesti:

> ***Quali tecniche basate su modelli di deep learning sono più efficaci per l'anonimizzazione automatica di dati clinici testuali, e in che modo BERT con Named Entity Recognition può essere utilizzato a questo scopo?***

Tra le tecniche di deep learning più efficaci per questo compito, i modelli basati su **transformer**, in particolare **BERT (Bidirectional Encoder Representations from Transformers)**, hanno

mostrato risultati all'avanguardia.

Modelli pre-addestrati su grandi corpora testuali, come BERT e le sue varianti biomedicali (ad esempio BioBERT o ClinicalBERT), si sono dimostrati particolarmente efficaci per il riconoscimento delle entità cliniche (NER), grazie alla loro capacità di catturare contesti bidirezionali e relazioni semantiche complesse. L'approccio NER consente di identificare e classificare automaticamente le entità sensibili presenti nei testi clinici, come nomi di pazienti, date, farmaci e condizioni mediche, rappresentando così una solida base per l'anonimizzazione dei dati.

Nel caso specifico dell'uso di BERT per l'anonimizzazione tramite NER, i risultati ottenuti mostrano un buon equilibrio tra recall e precision, con valori che nel nostro caso sono circa 0.81 per la precision e intorno a 0.88 per il recall. Questo indica una solida capacità di identificare entità sensibili, anche se con la presenza di alcuni falsi positivi. Tale bilanciamento è auspicabile poiché è preferibile individuare in eccesso entità non sensibili (falsi positivi) piuttosto che non anonimizzare dati realmente sensibili (falsi negativi), minimizzando così i rischi di esposizione di informazioni private. L'adozione di modelli BERT consente di contestualizzare efficacemente il testo e di risolvere ambiguità linguistiche, migliorando notevolmente la qualità dell'anonimizzazione rispetto a metodi basati su regole o modelli meno sofisticati.

4.3 Analisi dei risultati della Text Summarization

Dopo l'anonimizzazione dei testi clinici tramite BERT e mascheramento semantico, il flusso di lavoro prevede una fase di text summarization basata su ChatGPT e prompt engineering. In pratica, abbiamo chiesto a ChatGPT di generare un riassunto del testo anonimo, indicando esplicitamente di tenere in considerazione non solo il contenuto clinico, ma anche quali label NER risultassero più frequenti nell'intero insieme di trascritti, in modo da guidare il modello verso la preservazione delle entità più rilevanti.

C'è da considerare un aspetto importante: valutare la qualità di questi riassunti si rivela più complesso rispetto a un semplice compito di sintesi tradizionale: da un lato occorre misurare la **similarità semantica** tra il testo anonimizzato originale e il riassunto, dall'altro è necessario verificare quante delle entità sensibili o cliniche (le label NER) vengano effettivamente preservate.

4.3.1 BERTScore

BERTScore è una metrica di valutazione basata sui moderni modelli transformer (in particolare BERT) che confronta i vettori di embedding delle parole tra il testo di riferimento e il testo generato. A differenza di misure tradizionali basate su conteggi di n-grammi (ad esempio ROUGE), BERTScore cattura **somiglianze semantiche** profonde, riconoscendo sinonimi, parafrasi e varianti lessicali. Questo lo rende particolarmente adatto per valutare compiti di text summarization o generazione libera, dove l'obiettivo non è copiare parola per parola, ma preservare il significato. Nel nostro progetto di riassunti anonimizzati, BERTScore è fondamentale per quantificare in modo robusto quanto il riassunto mantenga concetti e relazioni chiave dal testo originario, pur consentendo variazioni linguistiche e la presenza di etichette di anonimizzazione.

Per la prima dimensione è stato quindi utilizzato **BERTScore**, ottenendo un punteggio di circa **0.80**, un valore che indica un buon livello di allineamento semantico nonostante la significativa

compressione del contenuto. Questa metrica ci ha permesso di quantificare quanto il riassunto mantenga la “sostanza” del testo di partenza, riconoscendo sinonimi e parafrasi.

4.3.2 Conteggio di etichette

Per la seconda dimensione, invece, è stato adottato un approccio numerico diretto: abbiamo contato quante label del testo originale comparissero nel riassunto e abbiamo riscontrato che **circa due terzi** delle etichette vengono mantenute. Questo binomio di misurazioni — una qualitativa (BERTScore) e una quantitativa (conteggio delle label) — garantisce una valutazione a tutto tondo del processo, bilanciando l’esigenza di sintesi e la necessità di completezza informativa.

Nel panorama più ampio delle **tecniche di text summarization** per l’anonimizzazione, si osserva una tendenza a combinare metodi estrattivi, che preservano parole chiave e frasi cruciali, con approcci astrattivi, capaci di generare testi fluidi e leggibili. Il **prompt engineering** gioca un ruolo cruciale: strutturando le istruzioni in modo da enfatizzare o escludere categorie cliniche specifiche (ad esempio farmaci, sintomi o procedure), si ottiene un miglior controllo sul contenuto prodotto dal modello. In particolare, l’aver integrato nel prompt l’informazione sulle label NER più frequenti ha consentito a ChatGPT di riconoscere immediatamente quali concetti fossero davvero centrali, migliorando sia la sicurezza (perché evita di reintrodurre dati sensibili) sia la qualità informativa del riassunto.

Infine, questa combinazione di **anonimizzazione tramite NER** e **text summarization basata su LLM** si configura oggi come uno degli approcci più promettenti in ambito di preservazione della privacy. La sua efficacia è confermata non solo dai nostri risultati sperimentali — con un BERTScore intorno a 0.80 e un tasso di preservazione delle label vicino al 66% — ma anche dalla letteratura recente, che sottolinea l’importanza di pipeline ibride in grado di mantenere il giusto equilibrio tra protezione dei dati e utilità clinica del testo sintetizzato.

4.3.3 RQ2

Grazie ai risultati ottenuti nel paragrafo precedente, è possibile rispondere alla seguente domanda:

> ***In che modo è possibile applicare tecniche di text summarization, basate su prompt engineering e arricchite con le informazioni semantiche estratte tramite NER, per generare riassunti anonimizzati che mantengano la significatività clinica dei testi originali?***

Per affrontare la seconda Research Question, abbiamo ideato una pipeline in cui la text summarization segue immediatamente l'anonimizzazione basata su NER, e viene resa possibile grazie a un accurato prompt engineering che sfrutta le informazioni semantiche estratte in precedenza.

In primo luogo, utilizzando BERT adattato al dominio biomedico, identifichiamo automaticamente le entità cliniche e sensibili, (come diagnosi, procedure terapeutiche, sintomi e valori di laboratorio) sostituendole con placeholder o pseudonimi per garantire la privacy. Successivamente, per la generazione del riassunto impieghiamo ChatGPT, inviando un prompt costruito su misura che invita il modello a concentrarsi sui contenuti clinici più frequenti e rilevanti emersi dalla fase NER, senza reintrodurre dati sensibili. In questo modo ChatGPT riesce a produrre un testo sintetico e privo di identificativi, ma con le entità chiave chiaramente rappresentate. La qualità dei riassunti viene quindi valutata secondo due criteri complementari: da un lato con BERTScore, che ci fornisce un F1 di circa 0.80 e conferma la conservazione delle relazioni semantiche principali; dall'altro contando quante delle label NER originali compaiono ancora nel riassunto, ottenendo circa due terzi di preservazione. Grazie a questa combinazione di anonimizzazione basata su NER, prompt engineering e metriche complementari (semantica e quantitativa) è possibile generare riassunti clinici che, pur rispettando i vincoli di privacy, mantengono pienamente la significatività e l'efficacia informativa del testo di partenza.

4.4 Conclusioni e sviluppi futuri

In conclusione, l'adozione di una pipeline ibrida basata sull'estrazione semantica tramite BERT e sul mascheramento controllato delle entità cliniche ha dimostrato di essere un prerequisito essenziale per poter applicare in sicurezza tecniche di text summarization con modelli LLM. In particolare, l'integrazione del prompt engineering, che ha guidato ChatGPT a concentrarsi sulle label NER più frequenti e rilevanti, si è tradotta in riassunti che, pur garantendo un elevato livello di privacy, mantengono circa il 66% delle etichette originali e raggiungono un BERTScore intorno a 0,80. Questi risultati confermano che è preferibile accettare un certo grado di falsi positivi nell'individuazione delle entità per evitare di tralasciare informazioni sensibili, ottenendo al contempo un testo sintetico, coerente e clinicamente significativo.

Come sviluppi futuri, risulta promettente arricchire il dataset con trascrizioni cliniche provenienti da diverse specializzazioni e strutture sanitarie, così da testare la robustezza della pipeline NER + summarization in contesti eterogenei. Un'altra direzione potrebbe riguardare il fine-tuning di ChatGPT o di modelli open-source similari direttamente sul dominio medico anonimizzato, allo scopo di aumentare ulteriormente precision e recall nella generazione dei riassunti. Infine, compiere simulazioni di attacchi di de-anonimizzazione potrebbe contribuire a mettere a punto un framework ancora più solido, capace di bilanciare al meglio le esigenze di protezione dei dati e di utilità informativa per applicazioni cliniche e di ricerca.